An Improved Classification Approach for Echocardiograms Embedding Temporal Information

Wei Li

June 2016

School of science and technology
Middlesex University London

A dissertation submitted to Middlesex University London in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Abstract

Cardiovascular disease is an umbrella term for all diseases of the heart. At present, computer-aided echocardiogram diagnosis is becoming increasingly beneficial. For echocardiography, different cardiac views can be acquired depending on the location and angulations of the ultrasound transducer. Hence, the automatic echocardiogram view classification is the first step for echocardiogram diagnosis, especially for computer-aided system and even for automatic diagnosis in the future. In addition, heart views classification makes it possible to label images especially for large-scale echo videos, provide a facility for database management and collection.

This thesis presents a framework for automatic cardiac viewpoints classification of echocardiogram video data. In this research, we aim to overcome the challenges facing this investigation while analyzing, recognizing and classifying echocardiogram videos from 3D (2D spatial and 1D temporal) space. Specifically, we extend 2D KAZE approach into 3D space for feature detection and propose a histogram of acceleration as feature descriptor. Subsequently, feature encoding follows before the application of SVM to classify echo videos.

In addition, comparison with the state of the art methodologies also takes place, including 2D SIFT, 3D SIFT, and optical flow technique to extract temporal information sustained in the video images.

As a result, the performance of 2D KAZE, 2D KAZE with Optical Flow, 3D KAZE, Optical Flow, 2D SIFT and 3D SIFT delivers accuracy rate of 89.4%, 84.3%, 87.9%, 79.4%, 83.8% and 73.8% respectively for the eight view classes of echo videos.

Acknowledgements

Firstly, I would like to thank my supervisor Prof. Xiaohong Wang Gao for providing this opportunity to study at Middlesex University in United Kingdom. She provided far more than just the required supervisory feedback, excellent guidance and constructive criticism throughout the journey of the research. I would also like to express my deepest gratitude to Prof. Martin Loomes and Prof. Qiang Lin for their thoughtful guidance and constant encouragement.

Many thanks go to the staff and research students at Middlesex University London for providing discussions, support and help with the program. Special thanks go to Dr. Yu Qian, Ammar Zayouna and Liqin Huang, who gave me a number of valuable comments about my research. And Lina Zhang, she gave me great help in friendship.

I would like to express sincere thanks to my mother-in-law and father-in-law for looking after my eight-year-old daughter and providing great support to me. I offer my gratitude to my husband, Chenhui Lin, whose inspiration prompted me following this academic journey. It is his sacrifices he made, unselfish supports and enduring encouragement over the years that I have been able to fulfil this accomplishment. All my love goes to my daughter, who supports me with her action, which makes me concentrate on my project.

Finally, this work is financially supported by the WIDTH project that is funded by EU FP7 Marie Curie programme and the School of Science and Technology, Middlesex University London. Their support is gratefully acknowledged.

Abbreviation:

A2C:	Apical 2 Chambers
A3C:	Apical 3 Chambers
A4C:	Apical 4 Chambers
A5C:	Apical 5 Chambers
AA:	Apical Angles
AAR:	Average Accuracy Rate
AER:	Average Error Rate
AML:	Anterior mitral leaflet
AO:	Aortic root
AOS:	Additive Operator Splitting
AoV:	Aortic Valve
AR:	Accuracy Rate
ASM:	Active Shape Model
BoF:	Bag-of-Features
BoVW:	Bag-of-Visual-Words
BoW:	Bag of Word
CMR:	Cardiovascular Magnetic Resonance
CT:	Computed Tomography
CVD:	Cardiovascular Disease
DICOM:	Digital Imaging and Communications in Medicine
DoG:	Difference of Gaussian
ECG:	Electrocardiogram
EF:	Ejection Fraction
ER:	Error Rate
FV:	Fisher Vector
GMM:	Gaussian Mixture Model
HOA:	Histograms of Acceleration
HOF:	Histogram of optical flow
HOG:	Histogram of Oriented Gradients
HOG3D:	Histograms of 3D gradient
IVS:	Interventricular Septum
LV:	Left Ventricle
LVOT:	Left ventricular Outflow Tract
mAP:	mean Average Precision
MRF:	Markov Random Field
MRI:	Magnetic Resonance Imaging
MV:	Mitral valve
PCA:	Principle component analysis
PLA:	Parasternal Long Axis
PM:	Papillary muscle
PMK:	Pyramid Matching Kernel
PML:	Posterior mitral leaflet
PMPM:	Postero-medical papillary muscle
PSA:	Parasternal Short Axis

PSAA: Parasternal Short Axis- Aorta
PSAB: Parasternal Short Axis-Basal
PSAM: Parasternal short axis-mitral
PSAP: Parasternal short axis- papillary
PV: Pulmonic valve
PW: Posterior wall
ROI: Region of interest
RVOT: Right ventricular outflow tract
SIFT: Scale Invariant Feature Transform
SPM: Spatial Pyramid Matching
SURF: Speeded Up Robust Features
SVM: Support Vector Machine
TTE: Transthoracic echocardiogram
TV: Tricuspid valve

Contents

Captions

List of Figures

List of Tables

1. Introduction

The human heart is a two-stage (systolic and diastolic) electrical pump that circulates blood throughout the body, the rhythm of which endows us with not only assurance of life but also a barometer indicating any potential abnormalities. Cardiovascular disease is an umbrella term for all diseases of the heart and circulation, including coronary heart disease, myocardial infarction stroke, atrial fibrillation, heart failure, cardiomyopathy and others. It is estimated that heart disease causes more than a quarter of all deaths in the UK, or around 160,000 deaths each year[1]. In the United States, about 610,000 people die of heart disease every year, which is 1 in every 4 deaths[2]. The situation in China is very severe as well. According official data[3], about 230 million people have cardiovascular disease, and projected annual cardiovascular events are predicted to increase by 50% between 2010 and 2030 based on population aging and growing alone in China. Therefore, improving the management of cardiovascular diseases is one of the greatest challenges faced by the healthcare industry in every country. So far many researches have been done for the diagnosis of cardiac diseases. Most of them are based on the analysis of cardiovascular imaging.

1.1 Cardiovascular Imaging

Cardiovascular Imaging has become the cornerstone of clinical diagnosis for cardiovascular disease (CVD). At present, there are three popular image-based test methods for cardiac clinical diagnosis including Echocardiography, Computed Tomography (CT) and Magnetic Resonance Imaging (MRI). Due to the advent of new technology and the refinement of existing technologies, imaging's role has

extended into biological, functional, and hemodynamic diagnosis of multiple pathophysiologic processes[4].

In clinical applications, Echocardiography has played a central role in the more comprehensive and reliable assessment of myocardial function (including systolic and diastolic function), according to [5–12]. On the other hand, Cardiovascular Magnetic Resonance (CMR) offers a comprehensive assessment of heart failure patients and is now the gold standard imaging technique to assess myocardial anatomy, regional and global function, and viability [13]. Cardiac MRI can provide a good assessment of cardiac function for some diseases, e.g. Chagas' disease ([14], [15]), congenital heart disease ([16], [17]), acute myocardial infarction ([18], [19]). In addition, an urgent CT of the thorax is frequently requested for suspected cases of aortic dissection or pulmonary embolism which may have atypical clinical features overlapping with those of acute myocardial infarction (MI)[20]. Multi-detector CT can provide complementary imaging of MRI, offering a combined morphological and angiographic assessment[21]. CT angiography can also assess ventricular systolic function, both global and regional, and myocardial perfusion[22]. Various imaging modalities can be used to assess cardiac structure and function, the presence and severity of dynamic obstruction, the presence of mitral valve abnormalities, and the severity of mitral regurgitation, as well as myocardial ischemia, fibrosis, and metabolic disease [23].

Among these modalities of cardiovascular imaging, echocardiography, applying ultrasound technique, has become routinely used in the diagnosis, management, and follow-up of patients with any suspected or known heart diseases. It is one of the most widely used imaging technologies in medicine[24]. It can provide a wealth of helpful information, including the size and shape of the heart (internal chamber size

quantification), pumping capacity, and the location and extent of any tissue damage. An echocardiogram can also give physicians other estimates of heart function such as cardiac output, ejection fraction (EF), and diastolic function.

1.2 Motivations and Objectives

In recent years, with the development of computer technology, considerable efforts have been made in Computer-Aided Diagnosis (CAD) using medical images to improve a clinician's confidence in analyzing medical images[25]. Evaluation of medical images by a clinician is qualitative in nature and may vary from person to person. A lot of research efforts have been directed to the field of medical image analysis with the aim to assist diagnosis and clinical studies[26]. When digital echocardiography is available, CAD may help physicians to make a more accurate decision [27]. At present, computer-aided echocardiogram diagnosis is becoming increasingly beneficial, this view is also shared in [28],[29].

For echocardiography, different cardiac views can be acquired depending on the location and angulations of the ultrasound transducer (to be detailed in Section 2.1). The major anatomical structures such as Left Ventricle (LV) are then manually delineated and measured from different view images to further analyse the function of the heart. Hence, the automatic echocardiogram view classification is the first step for echocardiogram diagnosis, especially for CAD system and even for automatic diagnosis in the future. For example, EF can be figured out by calculating the maximum and minimum diameter of LV cavity as discussed in[30]. Kumar, R. et al [29] propose that hypokinetic condition of the heart can be detected by extracting the feature points from the left ventricle wall and mitral valve leaflets. In addition, for the further analysis, collection and reduction of medical data and database management, automatic classification of heart views makes it possible to

label images or videos, providing a facility to manage all echocardiogram videos, while manual intervention based view recognition module can become a bottle-neck especially for large-scale echocardiogram images/videos. This view is also reflected in[31] and[32].

Image-based methods [31- 35] have become popular for echocardiogram classification owing to cardiac structural characteristics. The success of such approaches can be attributed to the structural dependence among cardiac cavities. However, there are still a number of challenges in improving classification performance. Firstly, when some cardiac structures change with different diseases or detection conditions, image-based matching, segmentation, location or other characteristic calculations can be affected by the high structural dependence. Secondly, most image-based methods focus on the integral shape variance of several particular cardiac structures, such as LV [36,37], but can pay less attention to some local inconspicuous structures, e.g. the aortic root in A5C viewpoint. Therefore the applicability of existing approaches for more cardiac viewpoints classification needs to be further evaluated.

To overcome these shortcomings, a number of progresses have been made. Some other methods [38, 39] proposed to recognize echocardiogram by fusing temporal information. The local motional structures or parts in the echocardiogram video are detected and represented in the spatial-temporal approaches. To combine both spatial and temporal information, the final classification results are usually obtained by using the voting scheme based on 2D images statistics.

In our research, we aim to respond to the challenges in a 3D space, i.e. 2D spatial and 1D temporal, to analyze and recognize echocardiogram videos with the following objectives.

- To treat each real video dataset as an entirety instead of dividing it into a set of frames;
- To focus on extracting 3D KAZE features;
- To propose acceleration field to describe features;
- To adopt Fisher vectors to represent feature descriptors;
- To improve the efficiency and accuracy of the classification of echocardiogram videos.

1.3 Main Contributions of the Thesis Proposal

The goal of this dissertation is to recognize the echocardiogram (echo) viewpoints automatically. In the process of each stage implementation, this work makes several contributions by showing the effectiveness and accuracy of the proposed methods for echo video classification, including

- In spatial-temporal domain, the 3D KAZE method is developed and implemented to detect features of the echocardiogram video. The method includes creating multiple non-linear scale spaces and using Hessian saliency measures to locate feature points. In order to exclude false feature points, we propose lower magnitude and neighbourhood suppressions to improve echocardiogram characteristics. It can also be used to detect 3D images with lower resolution.
- Based on the characteristics of myocardium motion, we introduce the acceleration field of cardiac motion as a novel features. Accordingly, histograms of acceleration descriptors are generated to describe the motion and stress state corresponding to particular viewpoints of the heart. Experimental evaluation demonstrates the promise of the proposed method for feature

description. It outperforms popular HOF descriptor without losing computational efficiency.

- We evaluate and compare the five existing local space-time feature descriptors for echocardiogram video recognition. We investigate their performance on a total of eight classes. On the base of 3D KAZE detector, all the descriptors are feasible for echocardiogram video classification, especially for spatial gradient descriptors, which can be affected by the noise of image more easily than the motional information descriptors.
- We modify BoW based on the standard baseline and apply it into encoding echocardiogram features. First, we choose several closer centres instead of only one when using k-means clustering. Next, we show how to incorporate spatial constrains in BoW models to improve accuracy for echocardiogram recognition. Additionally, the Fisher vector method is used to represent features as well. Finally, we evaluate BoW and FV performance on our dataset with varying parameters and reach a conclusion of applicability.
- On the base of the experiment, we measure the accuracy and efficiency and confirm an effective classification strategy when using SVM (Support vector machine) to recognize echocardiogram viewpoints.

1.4 Structure of the thesis

Since the classification of video images involves a number of stages, including feature point detection, representation, optimization, description and classification, this thesis is structured based on these stages. In particular, at each stage, comparison between our approach and a number of existing approaches is performed, analyzed and evaluated.

The dissertation begins with a literature review about the echocardiography and its classification in Chapter 2. All our datasets are illustrated in Chapter 3. Chapter 4 introduces our approach of 3D KAZE feature detection in echocardiogram videos whereas a number of results about our detectors are given in Chapter 5. In regard to the description method of echocardiogram features, the application of HOA (histogram of acceleration) descriptor, is proposed in Chapter 6. The resultant evaluations of the parameters and the comparison with a number of well-known descriptors are addressed as well in this Chapter. The results of echocardiogram classification based on 2D and 3D feature detection the comparison with other state-of-the-art are illustrated in the Chapter 7. In this chapter, video representation methods and the strategy when using SVM are proposed, which is followed by comparison results and strategic discussions. Finally, Chapter 8 summarizes the major contributions with a brief conclusion as well as a number of recommendations for future work.

2. Literature Review

The automatic identification of image content in medical images is an important pre-processing step in not only CAD (computer aided diagnostic) systems, but also in content-based image retrieval and picture archival systems[34]. CAD systems can calculate some structural parameters [40-42] or motion information [43] for further assessments on the base of medical images or videos corresponding to some particular tissue inspection.

Echocardiography is a typical application of Ultrasound imaging and is the most widely used tool in clinical practice for the evaluation of cardiac function due to its nature of easy to operate, inexpensive, non-invasive and in-vivo observation of the moving heart. As a result, the work for Computer-aided echocardiogram diagnosis become increasing. In doing so, the characteristic views corresponding to different transducer locations should be provided firstly. Unfortunately, echocardiogram classification is a considerably difficult work because of its native character. So, firstly we discuss the principles and characters that echocardiography sustains, and then investigate the current work in classifying echocardiogram views in comparison with normal action recognition.

2.1 Echocardiography

2.1.1 Introduction of echocardiography

Ultrasonic imaging is a mature medical technology. More than one out of every four medical diagnostic imaging studies in the world is now estimated to be an ultrasound study and this trend continues to increase[44]. This reason is mainly because of the remarkable advances that have taken place in the physics and

engineering of ultrasonic imaging since the medical applications of ultrasonic are introduced[45].

Echocardiography, also called echo, is the use of specialized ultrasound equipment to image the structure and function of the heart. It is rather like sonar, in that sound waves are used to locate the position of an object based on the characteristics of the reflected signal, hence the use of the term ‘echo’[46], which cannot be abbreviated as ECG that refers to an electrocardiogram. In acquiring a video clip, the echocardiography transducer (or probe) is placed on the chest wall surface (or thorax) of the subject, and images are then taken through the chest wall. This is a non-invasive, highly accurate and quick assessment of the overall healthy status of the heart[47]. A standard echocardiogram is also known as a transthoracic echocardiogram (TTE), or cardiac ultrasound. It has three basic "modes" that are used to image the heart: M-mode imaging, Doppler imaging and two-dimensional (2D) imaging.

The M-mode echo, which provides a 1D view, is used for fine measurements. Doppler mode ultrasound can be used to estimate the velocity of blood flow in the human heart and vasculature noninvasively[48]. 2D mode imaging is the mainstream of echo imaging and allows structures to be viewed in vivo in real time for any cross-section of the heart (two dimensions). In 2D mode imaging clips, all chambers and valves of the heart as well as the adjacent proximal connections of large vessels can be imaged and then spatial relationships among normal and abnormal intra-cardiac structures can be viewed. Cardiologists can assess motion characteristics of intra-cardiac structures in addition to their anatomy.

2.1.2 Physical and technical principles of ultrasound imaging

In order to accurately interpret echocardiogram, a basic understanding of the physical principles involved in ultrasound imaging is essential. Ultrasound is the term used to describe the sound of frequencies above 20 000 Hertz (Hz), beyond the range of human hearing[49]. And most cardiac applications are performed using frequencies from 2 million to 10 million hertz, or 2-10 megahertz (MHz) [48]. During the examination, sound travels in mechanical waves with a speed dependent on the density and elastic properties of the medium in which they are travelling[50]. When the sound wave, which is generated by electrical stimulation of a piezoelectric crystal in the transducer, is transmitted into the heart tissues, it is partially reflected back to the transducer from the layers between different cardiac tissues or scattered from smaller structures. The rest travels forward through the next tissues. The reflection depends on the variation of impedance of the two adjacent tissues, e.g. myocardium and blood. The transducer picks up echoes of the sound waves as they bounce off different parts of the heart and change into electrical pulses. These echoes are turned into moving pictures of the heart that can be seen on a video screen[51].

The screen image results from the information on strength, timing and position of the returning waves. The amplitude, or strength, of the returning echo wave is translated into the brightness (whiteness) of an echo pixel. These bright structures are termed as hyperechoic. In contrast, low-amplitude waves are translated into shades of grey-hypoechoic regions, while the structure without reflecting any waves corresponds to a black dot (anechoic). The vertical position of the echo pixel on the screen is based on the time delay between the emission and return of the ultrasound beam. Because velocity is assumed to be constant within soft tissue, quickly

returning echoes reflect superficial structures. Slowly returning echoes reflect deeper structures. Horizontal position of the echo pixel on the screen is based on the receiving piezoelectric crystal's location along the transducer[52].

2.1.3 Cardiac motion

The heart is a pump that creates the pressure that drives the flow of blood throughout the system. The heart motion includes two steps: systolic and diastolic. It can be viewed as the "motor engine" of our lives. Before illustrating the motion of the heart, we should introduce its structures briefly. As illustrated in[53], there are actually four chambers (spaces) inside the heart. Two top chambers are called atriums, while the bottom chambers correspond to ventricles. Each side of the heart forms its own pumping systems, a right heart and a left heart. Each half consists of an atrium and a ventricle. With these systems, blood always flows in only one direction because there are valves between atrium and ventricle. These valves open in one direction like trapdoors to let the blood pass through.

From the beginning of a cardiac circle, myocardium, electrically activated, begins to develop force and the accumulated force is translated into an increase in cavity pressure[54]. The process appears as the contraction of ventricle. With the left ventricle contraction strengthening, the pressure is large enough to open the aortic valve, and then the systemic loop begins. The same process corresponding to the pulmonary loop is beginning synchronously in right heart. Afterwards, the heart muscle relaxes, allowing the blood to flow and fill in the atria. When the atria are filled, the pressure opens the valves (including mitral valve and tricuspid valve) and the blood flows down into the ventricles. Then new round of contraction and relaxation starts with the following cardiac circle beginning. Under normal circumstances, each cycle takes 0.8 second[55].

2.1.4 Echocardiogram characteristics

Ultrasonic imaging has gained much popularity through medical sciences recently. Compared with other imaging modalities (such as MRI and CT), some of the main reasons of its prominence are due to high speed of imaging, portable instruments, as well as relatively inexpensive, non-invasive, displaying the image in real time, free of radiation risk and with the availability of pocket size [56, 57]. Furthermore, Ultrasound (US) images are tomographic, i.e., offering a "cross-sectional" view of anatomical structures, and can be acquired in vivo, thus providing instantaneous visual guidance for many interventional procedures. Besides these characters as mentioned above, echocardiography has unique ability in viewing the moving heart in real time and recording dynamic information of the cardiac structures in sequential (video) form, and thus providing an important diagnostic aid in cardiology for the morphological and functional assessment.

However, the quality of medical ultrasound image is generally limited due to the existence of noise owning to the loss of proper contact or air gap between the transducer probe and the human body[58]. The presence of this kind of speckle noise severely degrades the fine details and contrast resolution of the image, making it difficult to detect small and low contrast structures in body, as mentioned in [59,60,61]. In addition, the result of ultrasound imaging is easily affected by artificial factors, such as patient cooperation and physique, its relative dependence on a skilled operator, which can lead to different cardiac assessments.

2.1.5 Problem statement

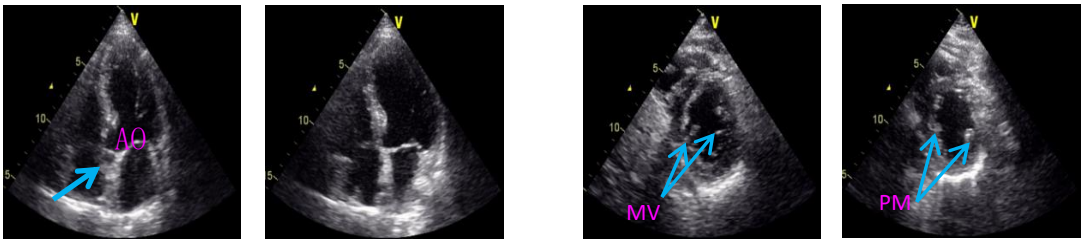
In order to identify cardiac structure automatically, as mentioned above, all of the previous work have worked to depict echocardiogram features in various forms and

to discriminate between images or videos by using many training and testing methods. The description and classification of cardiac feature are a very complex task [38,36] because of some following challenges exist:

(1) Intra-viewpoint variation. Because of the complicated cardiac motion (including systolic, diastolic, translation and rotation), speckle noise influence, patient individuality as well as instrument difference and the sonographer's experience, the image appearance belonging to the same cardiac viewpoint will present significant variations, which makes it difficult to achieve high classification accuracy (shown in Figure 2.1 (a)).

(2) Inter-viewpoint similarity. The high degree of structural similarity among the constellations of different viewpoints is another challenge. It is very difficult to discriminate those cardiac structural viewpoints with significant visual similarity by characterizing corresponding features. The ambiguities between similar viewpoints will exacerbate the situation with the availability of extra features replacing the missing structures to the sequence of echocardiogram as indicated in Figure 2.1 (b). The appearance of the left ventricle in PSAM and PSAP has significant similar because the characteristics between the mitral valve (MV) and papillary muscles (PM) are not prominent.

(3) Low quality of ultrasound image brings great difficulties in feature detection and description. How to highlight genuine features in echo image and restrain noise levels simultaneously is another critical factor for features representations.



(a) Two frames in a A5C video (b) PSAM (left) and PSAP (right)

Fig 2. 1 Some cardiac structures illustrating the variation of intra-viewpoint and inter-viewpoint. (a) The aortic root (AO) structure indicated in the left frame is missed in the right one; (b) The similarity between different viewpoints.

2. 2 Local Feature methods for video classification

A key component in any view classification systems is the representation – what feature set is used to represent a video? It is well-known in the pattern recognition community, that the proper choice of feature representation has a greater impact on performance, which is emphasized in [38] as well. In the following, we first discuss the existing feature detectors and feature descriptors, and then review Bag-of-Word model and Fisher vector methods respectively for feature representation.

2. 2.1 Feature detector

During the process of developing the CAD systems, feature extraction is one of the most important first steps for recognizing abnormal regions from the medical images[62]. Feature detectors tend to search some local regions and select characteristic locations and scales in videos by maximizing specific saliency response functions, which can be divided into two groups: 2D feature detection and 3D (spatial-temporal) feature detection.

2. 2.1.1 2D feature detector

For two dimensional images, Scale Invariant Feature Transform (SIFT) [63] is considered to be the more effective objection recognition technique for a variety of

applications. SIFT converts an image into a large collection of local feature vectors with invariant to rotation, scaling and translation. In[64], SIFT features and Pyramid Matching Kernel (PMK) are used to address object recognition. In[65], SIFT features combining with the local feature symmetry and the local energy (amplitude) filters based on a bag of visual words representation (detailed in section 2.2.3.1) are utilized to recognize anatomical structures of fetal heart and other tissues. SIFT builds a set of Gaussian scale spaces by using the Difference of Gaussians (mentioned in Appendix 1.6) operator and detects the maxima in each space as candidate key points. As depicted in [63], the final descriptor vector is with the dimension of 128 (an 4×4 array of histograms with 8 orientation bins in each). The experiment result obtained in [38] shows that a direct application of SIFT-based classification is ineffective in echocardiogram viewpoint discrimination problem.

In[66], another robust detector in the processing of image local feature detection is SURF (Speeded Up Robust Features), which is not only characterized by the invariance of scale and rotation but is also several times faster than SIFT. This method is based on the sum of 2D Haar wavelets technology (approximate Gaussian derivatives). It is applied to detect the key points of echocardiogram image in [35].

SIFT and SURF methods implement in a linear space, which smoothies to the same degree both details and noise at all linear scale levels. While nonlinear diffusion filtering make blurring locally adaptive to the image, so noise will be blurred while details will remain. For echocardiogram image, in order to retain the boundary and details of cardiac structures as well as to reduce noise level, KAZE feature method, which works in a nonlinear scale space proposed by Alcantarilla et al [67], has been applied to detect image features in echo classification by W. Li

et al. [68]. In nonlinear scale space, by computing the response of scale-normalized determinant of the Hessian matrix at multiple scale levels and searching for the maxima in both scale and neighbouring location, feature points can be detected. The rectangular grid around the detected feature points is divided into 4×4 subregions. The derivative responses $(d_x, d_{y,}|d_x|, |d_y|)$ in each subregion are weighted with a Gaussian function centred on the subregion centre and summed into a descriptor vector $V = (\sum d_x, \sum d_y, \sum|d_x|, \sum|d_y|)$. Accordingly, the final descriptor vector has a length of 64.

When using these three detecting methods to figure out the features of an echocardiogram image, it shows different results in highlighting the cardiac structural features. As evidenced in Figure 2.2, SIFT feature points (green points shown in Figure 2.2(b)) appear to spread across the entire image especially in the non-structure areas (e.g. inside of chambers), failing to highlight the structure of cardiac chambers boundaries, whereas SURF (corresponding to yellow points in Figure 2.2(c)) reduces points to a certain degree in the region of homogeneous areas, but contains some unreal features in chambers owing to its linear diffusion filtering which is similar with SIFT. However, the KAZE method (corresponding to pink points in Figure 2.2 (d)) improves the effect of noise reduction, and makes the cardiac chamber structure more outstanding. It shows better performance (comparing with SIFT and SURF features) in feature detection for echocardiogram images. The application of SIFT and KAZE feature detecting methods in echocardiogram recognition will be implemented in Section 7.3. All of these feature detectors are based on 2D spatial images. The 2D approach is limited as it can indicate the spatial information slice by slice, missing the temporal relationship between neighbouring slices in a video. As an extension study for echocardiogram

video classification, the spatial-temporal KAZE feature is the following focus in this research as illustrated in Chapter 4.

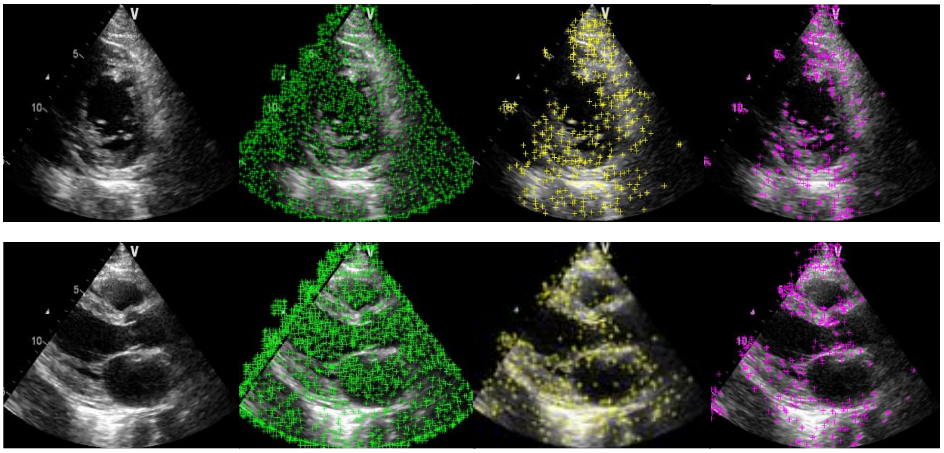


Fig 2. 2 The illustration of approaches of SIFT, SURF and KAZE on the extraction of feature points corresponding to LV (the top row) and PLA (the bottom row). (a) Original image of echocardiogram; (b) SIFT feature points (green); (c)SURF feature points (yellow) ; (d) KAZE feature points (pink) .

2. 2.1.2 3D feature detector

The extension of the SIFT approach to three dimensional data has been attempted by many researchers, including Ni et al. [69], Gu et al [70], and Allaire et al [71]. In 3D SIFT detection, multi-scale space is firstly defined as the convolution of the original 3D input volume with a variable scale of Gaussian function. Then DoG volume is formed based on this multi-scale space. Under each Gaussian scale, interest points are detected at the local maxima.

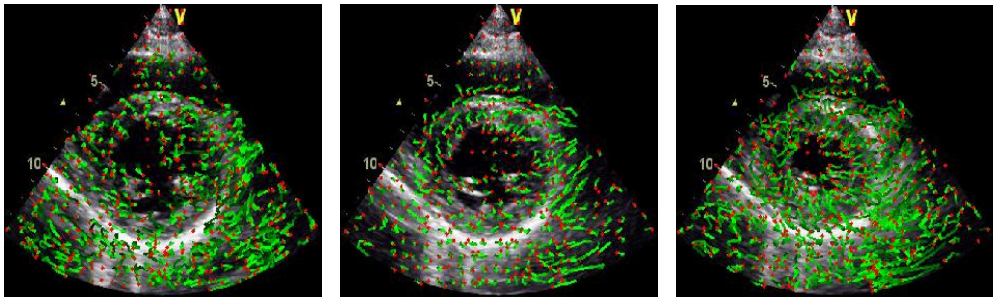
The Hessian3D detector, which is also called the 3D volumetric version of SURF by Yu et al [72] is proposed by Willems et al [73] as a spatial-temporal extension of the Hessian saliency measure applied for blob detection in images. They aim at a rather dense, scale-invariant, and computationally efficient interest point detector. The detector measures the saliency with the determinant of the 3D Hessian matrix. An integral video structure allows speeding up computations by approximating

derivatives with box-filter operations. A non-maximum suppression algorithm selects joint maxima over space and time scale.

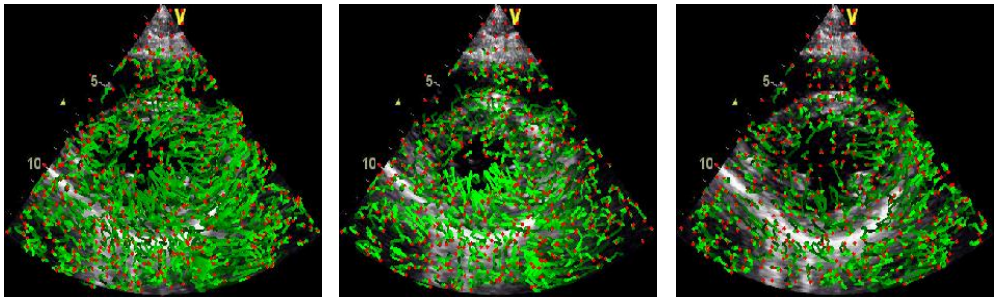
Laptev et al. [74,75] appear to be the first to propose a feature detector based on a spatial-temporal extension of the Harris corner criterion [76]. This Harris3D method is based on the eigenvalues of a spatial-temporal second-moment matrix at each video point. As a result, local maxima indicate feature points. Concerning the Harris criterion in Harris3D method, Dollár et al [77] considered that true spatial-temporal corner points are relatively rare in general cases. They therefore design their interest point detector to yield denser coverage in videos and employ spatial Gaussian kernels and temporal Gabor filters. As a result, local maxima give final interesting positions. The similar detector is utilized in [78] for echocardiogram video classification.

Feature trajectory, which differs from space-time point detectors as mentioned above, is a straightforward concept that reflects spatial-temporal information in video sequences. Trajectories corresponding to spatial interest points are tracked based on KLT tracker initiated by Lucas and Kanade [79] or dense optical flow field proposed by Lu et al. [80], whereby the flow shape encodes information about local motion patterns and can thus be directly used as local features. Matikainen et al. [81] track spatial interest points by using a standard KLT tracker, and then form a set of fixed length feature trajectories for the subsequently action recognition. Wang et al. [82] densely sample feature points at several spatial scales. Tracking these points is achieved by median filtering in a dense optical flow field[83]. This method is applied to depict the feature trajectories of echocardiogram video sequences where the result is illustrated in Figure 2.3. Dense sampling ensures a

good coverage of the video with features, but for echo image, many unintended points (speckle noises) are tracked.



Dense trajectory in cardiac systolic



Dense trajectory in cardiac diastolic

Fig 2.3 Visualization of Dense trajectory for the systolic and diastolic action in echocardiogram PSAP video sequence: in cardiac systolic, the LV begins to contract and the feature trajectories demonstrate the toward-center motion state and vice versa.

2. 2.2 Local feature descriptor in details

For each detected feature point (x, y, t), the feature descriptor should be computed to represent the characteristics of the local area corresponding to the feature point. A key factor for high accuracy classification is the use of local descriptors, which express integral information of features. In this research, how to combine dynamic information (just like histogram of optical flow and motion boundary histogram in some literatures) with the spatial state is the focus for constructing spatial-temporal descriptor.

The descriptors for local information can be divided in two classes, which are local appearance and motion information descriptors. The HOG (histogram of

gradient) is the most popular method for local appearance representation, which is widely used in action recognition. While for 3D appearance descriptor, 3DSIFT as well as HOG3D are representative in encoding local features. All of these descriptors are focusing on the static appearance information of the video sequence. Considering the dynamic information as mentioned above, HOF (histogram of optical flow) and MBH (motion boundary histogram) are proposed to describe the motion state in video classification, which can be called motion information descriptor.

Among the first works on local descriptors for videos, Laptev and Lindeberg [84] defined and compared several descriptors over local spatial-temporal neighbourhoods including single- and multi-scale N-jet derivatives, histograms of spatial-temporal gradients, histograms of optical flow, principle component analysis (PCA) of spatial-temporal gradient and PCA of optical flow. Among histogram methods, a feature vector is obtained by accumulating components of each cell with a certain size around the interest points. A different feature vector consists of projections of local image measurements onto D principle components by applying PCA to highlight the most significant eigenvalues. In their experiment, the histograms based on optical flow and spatial-temporal gradients show better performance than others. In their later work, Laptev et al. [85] introduced the histogram of oriented gradient (HOG) and the histogram of optical flow (HOF) descriptors to characterize the local information. The coarse HOG and HOF are computed in each cuboid divided from each space-time volume. Finally, the normalized cuboid histograms are concatenated into the HOG/HOF descriptor. In the work described by Wang et al [82] and Uijlings [86], motion boundary

histogram (MBH) combining with HOG and HOF is introduced for action recognition.

HOG descriptor is introduced by Dalal and Triggs in [87] for human detection, which can be used to reflect static appearance information. For the digital image, each pixel of the gradient image measures the change in intensity of that same point in the original image, in a given direction. In implementation, gradient magnitude responses are calculated in horizontal and vertical directions. The gradient of the image is computed using the finite difference approximations:

$$L_x(i,j,t) = (I(i+1,j,t) - I(i-1,j,t))/2$$
$$L_y(i,j,t) = (I(i,j+1,t) - I(i,j-1,t))/2 \quad (2\text{-}1)$$

Where I is the intensity of the corresponding pixel. In order to reduce the noise effects, a video sequence can be filtered by using Gaussian operator or others. The magnitude and orientation are computed from the intensity gradients for every pixel in the gradient image. The orientation θ is in spatial domain with the range of $[-\pi, \pi]$. As reported by Uijlings et al [86], they divide θ into equally sized bins, and vote the weighting based magnitudes into the corresponding bins.

3D SIFT descriptor, as an extension of the SIFT descriptor [63], is proposed by Scovanner et al.[88], as shown in Figure 2.4. For a set of randomly sampled interest points, spatial-temporal gradients are computed for each pixel. Each pixel around the interest points is weighted by a Gaussian filter centred on the given position. For orientation quantization, two ways are provided: one way is to split both azimuth and elevation (φ, θ) into 8×4 bins using meridians and parallels; the second way is to tessellate the sphere using an icosahedron, which is a similar idea to the HOG3D descriptor. It increases the dimensionality to 2048 in [88], which put more constraints on efficiency. In [78], the polyhedron (with 80 triangle faces)

replaces the spherical coordinates (with 8×4 bins) to quantize the orientations of 3D SIFT descriptor in the echocardiogram video description. The neighbourhood of a feature point is selected with the size of $12 \times 12 \times 12$ and then divided into $2 \times 2 \times 2$ sub-volumes. The final descriptor is of 640 ($2 \times 2 \times 2 \times 80$) dimensions.

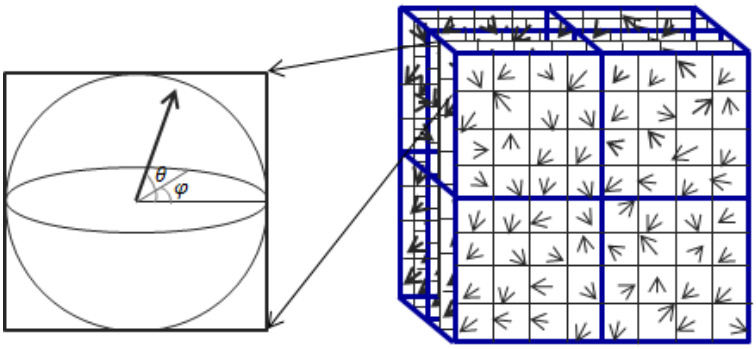


Fig 2. 4 The illustration of 3DSIFT descriptor using the first orientation quantization. 3D sub-volumes are sampled surrounding the interest point. Each sub-volume is accumulated into its own sub-histogram. All of these sub-histograms are concatenated into the final descriptor.

HOG3D descriptor, similar to 3D SIFT descriptor, is proposed originally in [89]. It is based on histograms of 3D gradient orientations and can be seen as an extension of SIFT descriptor [63] for videos. For a given cell, they divide it into $S \times S \times S$ sub-blocks, as shown in Figure 2.5 (b). The mean 3D gradient of each sub-block is then computed by using an integral video and subsequently quantized by employing a regular polyhedron (shown in Figure 2.5 (c)). The histogram for the given cell is obtained by summing the quantized mean gradients of all sub-blocks. All 3D gradient histograms corresponding to $M \times M \times N$ cells are finally concatenated to one feature vector. In [89], a local support region around a feature point is divided into a set of $4 \times 4 \times 3$ cells. All histograms quantized using an icosahedron for all are concatenated to form the final vector for the corresponding local region. This descriptor has higher dimensionality with 960 ($= 4 \times 4 \times 3 \times 20$).

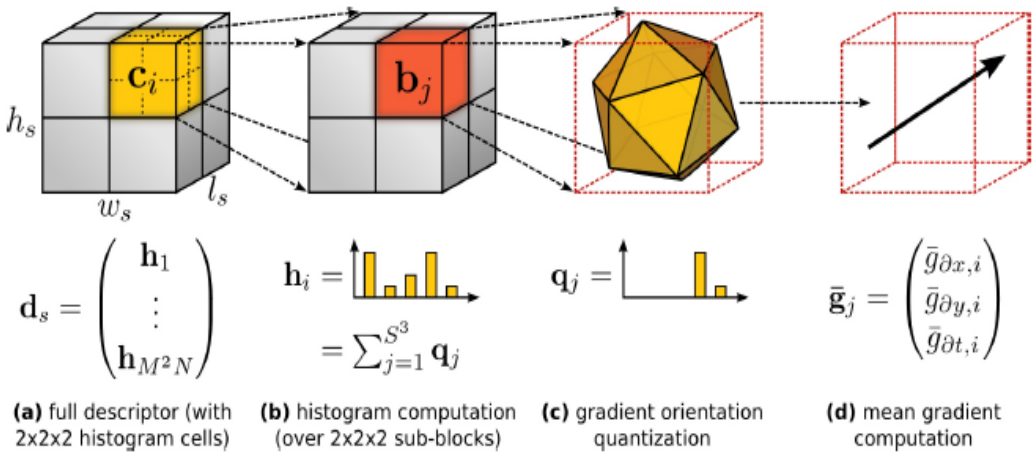


Fig 2. 5 The illustration of HOG3D descriptor (reprinted from [89]). Its orientation quantization method using regular polyhedron is similar with the second way depicting in 3D SIFT descriptor.

HOF descriptor captures the local motion information of the videos. Optical flow is the pattern of apparent motion of objects, surfaces, and edges in a visual scene, which includes two components (horizontal and vertical). These two components (v_x, v_y) are employed to compute orientation and magnitude of the motion. In our experiment, the classical Horn-Schunk [90] method is applied to calculate the dense optical flow responses. We use the version implemented by the Matlab software (www.mathworks.com). The orientation and magnitude of the optical flow are used as weighted votes into local orientated histograms in the same way as for the standard HOG.

MBH descriptor is another method to represent motion information on action recognition, which is introduced by Dalal et al. in [91], which has been combined with other descriptors in [82] to encode the local feature characteristics of human action. It is obtained by computing derivatives separately for the horizontal and vertical components of the optical flow. So it can be viewed as the gradient of the

optical flow. The orientation information is quantized into histograms, and the magnitude is used as weighted votes into local oriented histograms.

In practical applications, HOG3D and 3D SIFT descriptors introduce the third temporal gradient into the gradient component and form the 3D gradient as (L_x, L_y, L_t) , which causes higher dimensionality and hence reduces efficiency in the following classification task. Comparatively speaking, HOG is very efficient to compute with lower dimensions. However, it only encodes the appearance of local features and lacks of motion information for video recognition. Therefore, HOG descriptor often combines with HOF or MBH descriptor to reflect both appearance and motion details, as reported in [85,82].

Willems et al. [73] have proposed the extended SURF (ESURF) descriptor which extends the image SURF descriptor to videos. Similar to 3DSIFT, a rectangular volume defined around a point of interest is subsequently divided into a set of $M \times M \times M$ cells. Each cell is represented by a weighted sums ($v = (\sum d_x , \sum d_y , \sum d_t)$) of uniformly sampled responses of Haar-wavelets along the three axes ($d_{x,} d_y, d_t$).

2. 2.3 Feature Representation

A general approach to describe an image for classification is to extract a set of local features and subsequently descriptors, and represent them into a high dimensional vector. By far, the main paradigm in image representation is the well-known Bag-of-Words (BoW) model introduced by Sivic and Zisserman [92] as well as Fisher vector (FV) by Sánchez J et al.[93], and Simonyan K et al. [94].

2.2.3.1 Bag-of-Words (BoW)

BoW sometimes called Bag-of-Visual-Words (BoVW) or Bag-of-Features (BoF) in computer vision, can be applied to image classification by treating image features

as words (detailed in Appendix 1.4). Initiated for text classifications, Niebles [95], Wang et al. [96], Bilinski et al [97]extend this model into videos. For BoW representation in videos, visual vocabulary (or codebook) is produced by clustering all vectors obtained from training videos. K-mean is generally applied to generate k clusters which are referred to as visual words (or words). Video sequences are then represented as occurrence histogram of visual words. Yang et al. [98] propose a sparse coding algorithm to replace K-means clustering and a max-pooling of the descriptor-level statistics. Yu Qian et al. [78] employ this method instead of k-means to train an echocardiogram video vocabulary.

2.2.3.2 Fisher Vector (FV)

Fisher vector essentially is an extension of BoW developed by Lei et al. [99], which shows an improved performance over BoW for both image and action classification in the work by Chatfield [100], Sánchez J et al [93], Oneata et al. [101], and Kantorov et al [102]. In a nutshell, FV encoding aggregates a large set of feature descriptors into a high-dimensional vector representation. It is completed by fitting a parametric generative model, e.g. the Gaussian Mixture Model (GMM), to the features, and then encoding the derivatives of the log-likelihood of the model with respect to their parameters as discussed by Jaakkola et al. [103]. For a descriptor with size D, the GMM model with K Gaussians is first learned based on the training datasets. Then the first and second order derivatives corresponding to the given Gaussian mean and standard deviation with D-dimension are generated. The final FV is the concentration of the two parts and is therefore $2DK$-dimension as denoted by F. Perronnin et al. [104]. Sánchez J et al. [93] considers that the BoW is a particular case of the FV where the gradient computation is restricted to the mixture weight parameters of the GMM. Their experiment shows that the additional

parameters (including mean and deviation) incorporated in the FV bring large improvements in terms of accuracy for image classification. And it can be computed from much smaller vocabularies with K Gaussians and therefore at a lower computational cost.

2. 3 Related works for echocardiogram classification

Numerous works and methods have been proposed in the past within the field of automatic classification of echocardiogram images/videos. It is a challenging work because of not only the complicated characters of ultrasound image but also the variations of intra- and inter-viewpoint of echo videos. In addition, there are several viewpoints according to the different cardiac views with great structural similarity.

With regards to previous works in echocardiogram classification, there are two primary trends to detect and descript echo cardiac viewpoints, including image-based method and spatial-temporal fusion technique. This section reviews the state-of-the-art methods for echo video recognition based on these two categories, including

- *Image-based methods* (to be detailed in Section 2.3.1) focus on spatial relationship of the heart structures, and cardiac viewpoint recognition is carried out by using information of structural location, intensity as well as a number of related statistical characters.
- *Spatial-temporal fusion methods* (to be addressed in Section 2.3.2) explore spatial-temporal information of the heart structures to discriminate echocardiogram viewpoints by combining the motion information of echocardiogram sequences with the spatial and textural information.

2.3.1 Image-based methods

Image-based methods recognize echocardiogram images by employing spatial information of cardiac structures such as location, gradient, energy and other statistical characters. Approaches in this field are focusing on mining and collecting spatial features or relationship of cardiac structures and representing these features into other forms, and finally training and testing by using classification methods. Because temporal information is omitted in these approaches, the whole echo cardiac video can be divided frame by frame, and features and spatial relationships are extracted from each image.

Balaji et al. [37] extracted the diastolic frames from the echocardiogram videos as input images. After reducing noise and enhancing the contrast of the image, they utilize morphological grayscale closing to highlight the cardiac cavity, and then using connected components labelling to segment chambers. In their experiment, three standard views PSA (parasternal short axis), A2C and A4C can be classified.

Another work that models the structure of the heart is proposed by Ebadollahi et al [33]. The original cavities of the heart are detected by applying Grey-Level Symmetric Axis Transform (GSAT). The constellation of chambers in each image is viewed as a relational structure with some attributes (e.g. location, area and directionality of each part, and distance and angle between a pair of parts) (as shown in Figure 2.6).

The statistical variations and spatial relationships of the constellation (as the blue blobs and the red lines) are encoded by using Markov Random Field (MRF) models, and the optimal energy can be obtained subsequently. Final classification is then conducted by applying support vector machine (SVM) in energy space. In implementation, the process of detecting chamber varies to a certain extent, which

is sensitive to noise in ultrasound where small intensity variations can lead to large errors in a graph formation.

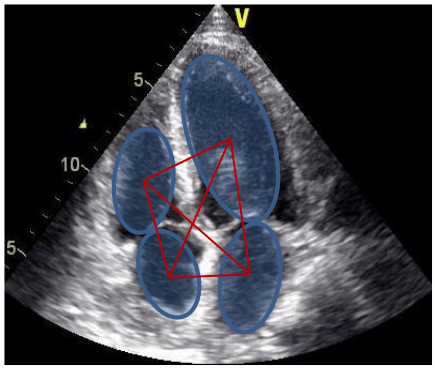
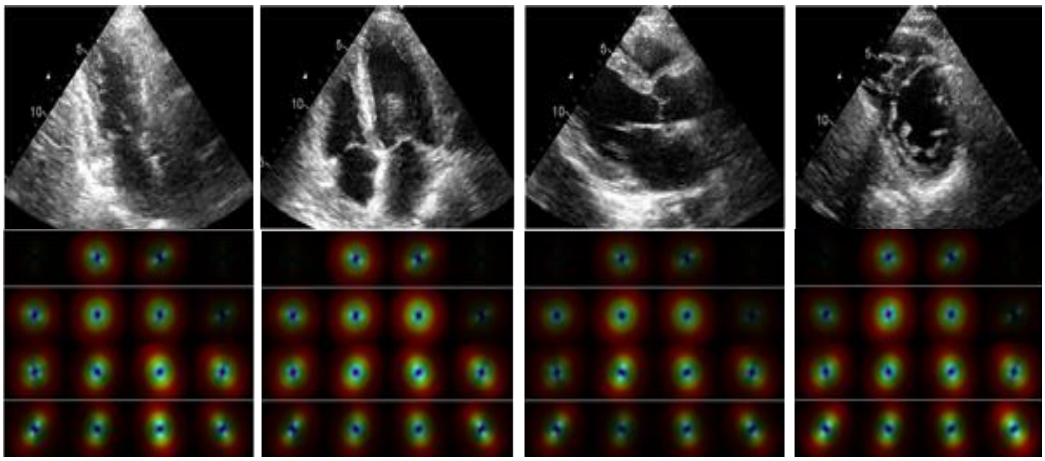


Fig 2. 6 The relational structure of cardiac cavities with the constellation of chambers (the blue blobs) and the relationships (the red lines) between a pair of chambers in A4C viewpoint.

In[34], the standard views of cardiac echo image are represented as template library by applying multi-resolution spline filters to intensity images. For each unknown sample image, the deformation map and warped image corresponding to each template can be obtained by matching it against the template library. Then both deformation energy and the similarity between the warped image and reference templates are used to classify the sampling image by applying a linear discriminant classifier. Such multi-scale elastic deformation is effective only when the unknown image viewpoint is close to known viewpoint templates with smaller deformation energy and more similarity.

Hui Wu et al. [105] divided the echocardiogram image into a set of non-overlapping image blocks and computed the spectral energy of each block by using the GIST feature method on the base of multiple oriented Gabor filters of different scales, which can be illustrated in Figure 2.7. Then, the features from each block are concatenated into a single feature vector to represent the global information of

the echocardiogram image. The final classification results are obtained after training by using SVM.



(a) A2C (b) A4C (c) PLA (d) PSAM

Fig 2. 7 The GIST features (the bottom row) represented in spectral energy reflects the global structural information of the corresponding different viewpoints (the top row)

In [36], a comprehensive system for cardiac echo video classification is described by using MLBoost Learning algorithm along with multi-object detection and integrated local-global features. On the base of developing the Left Ventricle (LV) detector by using Haar-wavelet type local features, each view can be represented by the global template based on the spatial layout of LV structure in other cardiac chambers. Given an input sequence, LV structure can be detected by applying the LV detectors. The corresponding global templates are constructed and fed into multi-view classifiers. The similar method using Haar-like local rectangle features and a multiclass classifier is also mentioned in [106]. Their work collects both positive and negative training images corresponding to all chambers and background respectively by encoding the structures into different template designs. On the base of multiclass classifier, the final classification result can be determined according to the majority voting rule.

Artificial Neural Networks (ANNs) are one of the popular methods for classifications, which have been of increasing interest in medical image processing (such as in [107], [26], [25]). Elalfi et al. [25] pre-process medical echocardiography images using Gaussian and Gabor filters and combine intensity histogram features and Gray Level Co-occurrence Matrix (GLCM) features, and then input these global texture features into an artificial neural network for automatic classification based on back-propagation algorithm to classify heart valve diseases.

In spatial feature detection for echocardiogram image, the idea of statistical histogram is proposed frequently. As one of the works in this direction, Roy et al. [108] have reasoned that the number of cavities that is present, their orientations, and the presence of heart muscles in each view generating different histograms. They propose the use of simple gray-scale (intensity) histograms in a region of interest (ROI) for four views classification. The final classification is made by using a neural network classifier where the number of hidden layer units is empirically chosen. The signature histogram for a given echo image is heavily dependent on ROI for which intensity values are considered, and the choice of this region is not made explicitly in this work. The similar method is proposed by Balaji et al. [109]. The histogram feature of gray scale is provided to characterize the echocardiogram image and then combined with SVM and Back Propagation Neural Network (BPNN) by Rumelhart et al. 1986 [110] to classify four views of echocardiogram. Histogram of Oriented Gradients (HOG) feature method is also applied to depict the spatial arrangement of echocardiogram image by Agarwal et al. [31]. It captures local structure while not requiring explicit correspondences to match images. Firstly, the sectorial portion of the ultrasound image is transformed to its rectangular beam

space equivalent. Then a set of local histograms corresponding to all rectangular cells are concatenated into the HOG feature vector of the image. View classification for PLA and PSA is accomplished by using SVM classifier.

In [34], Speed Up Robust Features (SURF) [111] are applied to detect and descript the key points of echocardiogram image, as shown in Fig 2.8. In the process of classification, the input echocardiogram images with extracted SURF descriptors need to match the descriptors of template image pre-computed in each category. The classification results can be obtained by calculating the minimal Euclidean distance between the matched points in the templates and the input echocardiogram image. The accuracy of recognition is highly dependent on the matching degree.

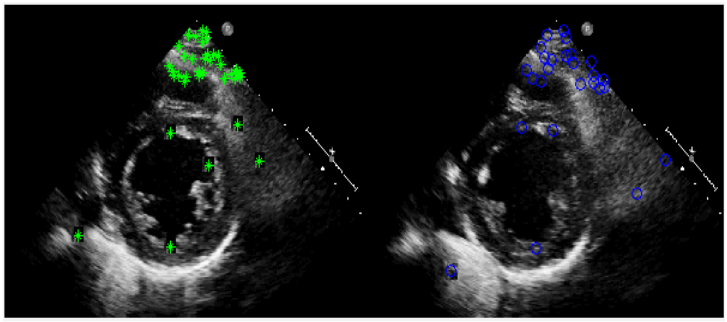


Fig.2.8 The matching points between template image and test image; (left) template image with SURF feature points (green); (right) input test image with matching points(blue) (courtesy of [35]).

A hierarchical classification strategy for view classification is proposed by Otey et al. [112]. They combine the principle features with other features (including gradient features, peak features and other statistical features) to form a feature vector corresponding to each image, and then use hierarchical classifiers (one on the top level, two classifiers on the second level) to classify all the images by using leave-one-out cross-validation rule. The top level classifier is used to distinguish between the apical and parasternal views, while on the second level, one is applied to distinguish two or four apical chamber viewpoints, and another is used for

parasternal long or short viewpoint recognition. This method implements the echocardiogram classification following top-down process, i.e. from two viewpoints to four viewpoints.

2.3.2 Spatial-temporal fusion methods

Compared with the image-based methods, temporal or motion information is also important while combining with spatial features as mentioned in image-based methods. The difference between the two kinds of approaches is that some local motional structures or parts in echocardiogram video are detected and represented in the spatial-temporal method, while global information or dense features are depicted in most image-based methods. Local space-time features capture cardiac characteristic appearance and motion information for a local region and provide a relatively independent representation of structures with respect to their spatial-temporal shifts and multiple motions in the scene. Such features are usually extracted directly from videos and therefore avoid possible failures of other pre-processing methods such as segmentation in [37] and matching in [34, 36]

In [38], local spatial-temporal features for each image are detected and encoded to recognize the echocardiogram images. In their framework, on the base of optical flow field, the edge-filtered motion maps for echocardiogram video sequences are generated by filtering the motion magnitude images based on edge maps, as shown in Figure 2.9. Local spatial-temporal features are detected using scale-invariant features [63] and then described by concatenating location, motion histogram and intensity histogram into a feature vector. In training process, a hierarchical dictionary and parameters of a Pyramid Matching Kernel based SVM[113] is learnt from all the training frames. Each frame of echocardiogram video is individually classified and final classification is achieved by using the voting scheme.

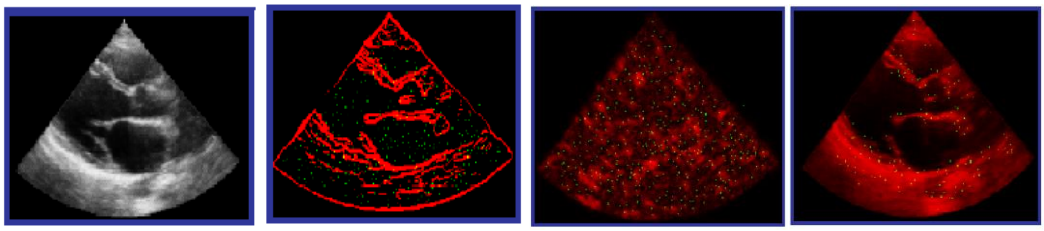


Fig 2. 9 The process of locating the structural feature points (courtesy of [38]).

Beymer et al. [39] proposed to exploit the motion information present in the echocardiogram videos for view classification. They use Active Shape Models (ASMs) to capture the shape and texture information and then track ASMs across the whole sequence to derive motion information. All the information is combined by projecting it down to a low variance of Eigen-motion feature spaces and the final classification is done by minimizing a ”sequence fit” measure. One of the limitations about this method is that original ASM feature points need to be located manually in the training data, which can be not only time consuming, but affected by human factors.

The space-time interest points of echo video clip are detected by applying Cuboid detector (including the 2D spatial Gaussian smoothing kernel and 1D temporal Gabor filter) by Qian et al. [78]. In this method, each interest point is represented into a 640-dimention vector by using 3D SIFT descriptor. In training process, a codebook of echocardiogram videos is constructed by following the Bag of Word (BoW) paradigm. Then all 3D SIFT features corresponding to space-time interest points in each testing videos are coded into a feature vector on the base of trained codebook. Multiclass SVM is applied to complete eight viewpoints classification of echocardiogram video in the final experiment.

Although spatial-temporal fusion methods have shown feasibility for multi-viewpoint recognition in non-normalized echocardiogram video data as mentioned above, correlational discussions are very limited in comparison with image-based methods. In addition, the classification accuracy rate of 72% in [78] appears to be in need of improvement. Therefore, the research about space-time features detection and description for echocardiogram video need to be studied further.

3. Datasets and Viewpoint Definition

In this section, we present the datasets that are used in our research. A total of 432 echocardiogram videos are collected from 93 different patients of 7 to 85 years of age (containing 35 wall motion abnormalities and 58 normal cases) in the hospital of both Tsinghua University Hospital and Fuzhou Hospital in China. All videos are captured from GE Vivid 7 or E9 and are stored in DICOM (Digital Imaging and Communications in Medicine) format with the size of 341 × 415 pixel × 26 frame.

3.1 Echocardiogram viewpoint definition

During an echocardiogram scanning, a sonographer images the heart by using an ultrasound machine/scanner and a transducer is placed against a patient's chest. Reflected sound waves reveal the inner structure of the heart walls and the velocities of blood flows. Since these measurements (e.g. heart and blood vessel morphology, heart wall thickness, blood flow velocity and so on) are typically obtained by using 2D images of the heart, the transducer position can be changed during an echo exam which captures different anatomical sections of the heart from different views, as shown in Figure 3.1.

The most common cross-sectional views of echocardiogram are the parasternal long axis (PLA), the parasternal short axis (PSA), and the apical angle views (AA) [114]. By varying the orientation of ultrasonic transducer respectively on these different views, sonographers can detect the cardiac structures and their motion status of different cardiac cycle at different anatomical sections from multiple angles, e.g. multiple chamber viewpoints.

In this dissertation, the term 'view' corresponds to the macroscopical cardiac tissue in echocardiogram (e.g. parasternal view or apical view) whereby a

transducer is positioned physically at one location for each view. While 'viewpoint' means the detailed structure showing in each view position. For example, there are a viewpoint of two chambers and a viewpoint of four chambers in the apical view, where the transducer changes incident angles at the same spot.

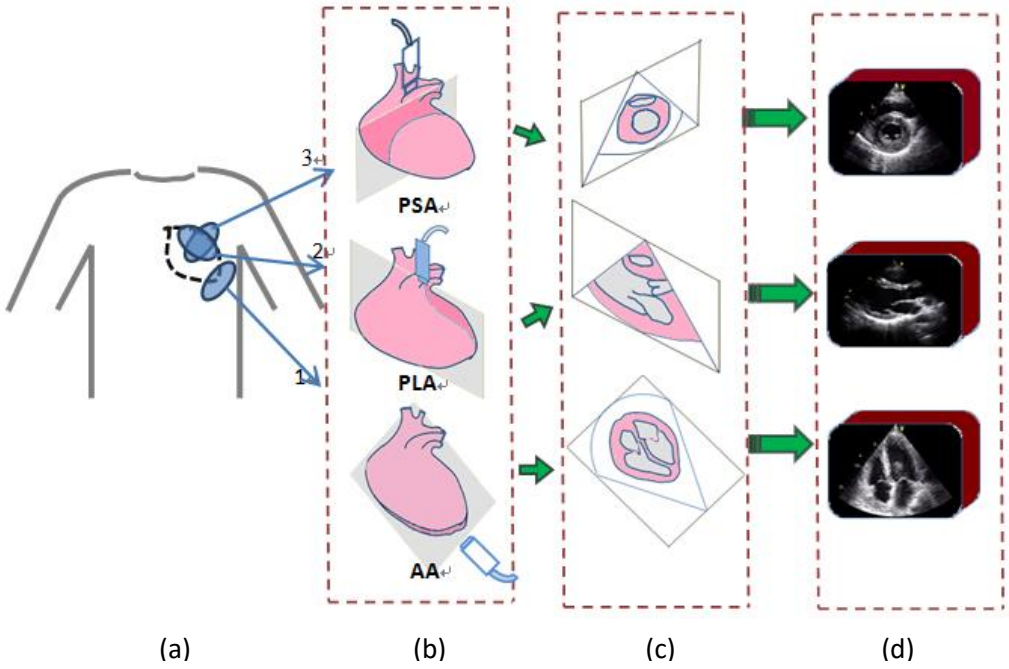


Fig 3. 1 The process of echocardiographic imaging. (a) The location of the three primary echocardiographic views on the chest ; (b) The planes that the echo beams scan from the transducer; (c) Cross-section drawn along the echo beams; (d) The cardiac structure image/video displaying in the screen.

Concretely speaking, there are three levels sampling the heart in PSA, including the level of the aortic valve, level of the mitral valve and papillary muscles. From the visual point of view, these three levels are called three viewpoints in this research which include parasternal short axis-aortic valve (PSAA), parasternal short axis-mitral valve (PSAM) and parasternal short axis-papillary viewpoint (PSAP). In a similar way, there are four viewpoints for AA, i.e. apical two chambers (A2C), apical three chambers (A3C), apical four chambers (A4C) and apical five chambers (A5C). So, there are altogether eight viewpoints including one parasternal long axis

(PLA) viewpoint. Table 3.1 shows the correlation between three views and corresponding viewpoints. Together, all the videos corresponding to every viewpoint make up the dataset used in the following research, the number detail is stated in Table 3.2.

Primary views	Primary view 1 (AA)	Primary view 2 (PLA)	Primary view 3 (PSA)
Sub-views	Viewpoint 1: A2C	Viewpoint 5:PLA	Viewpoint 6:PSAA
	Viewpoint 2: A3C		Viewpoint 7:PSAM
	Viewpoint 3: A4C		Viewpoint 8:PSAP
	Viewpoint 4: A5C		

Table 3. 1 Three primary views and eight viewpoints

Viewpoint	A2C	A3C	A4C	A5C	PLA	PSAA	PSAB	PSAP	Total
Video	62	46	58	40	79	57	48	42	432

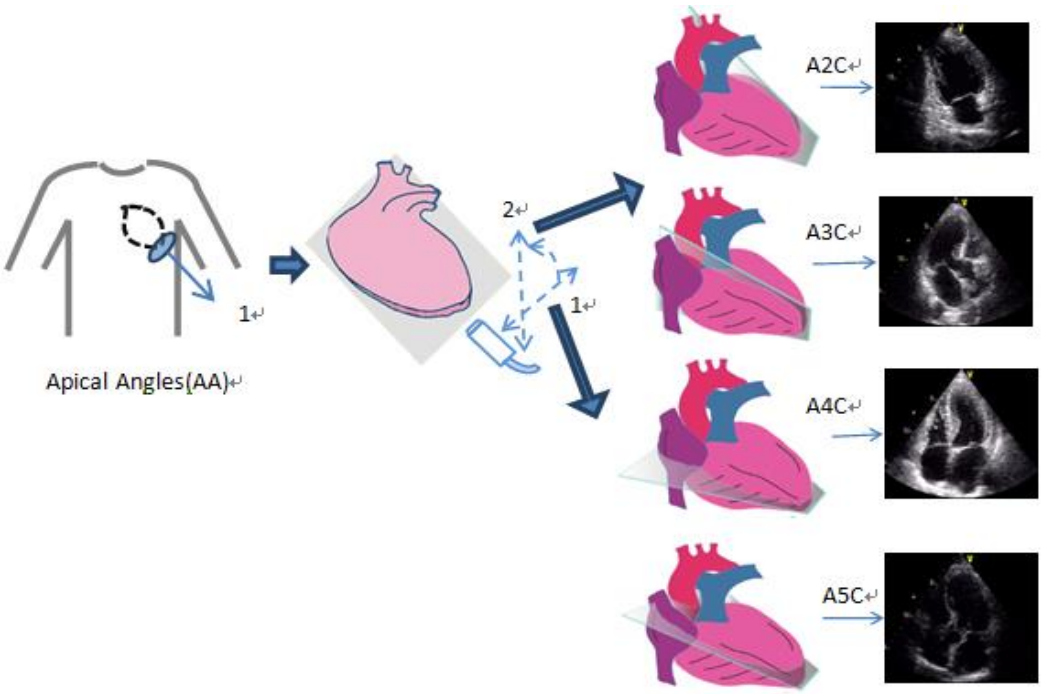
Table 3. 2 Eight viewpoints in the dataset

3.2 Datasets

3.2.1 Apical view

For apical view, four viewpoints are detected and displayed according to different chamber structures such as A2C, A3C A4C and A5C. The apical viewpoints are obtained by placing the probe at the point of apical impulse, which is located between the ribs on body chest, as shown in Figure 3.2 (a). The A4C is usually obtained first by the probe indicator (notch) at 3 o'clock (the orientation 1 in Figure

3.2 (b)), and the A5C can be detected by tilting the head of the probe upwards slightly along this direction. By rotating the probe approximately 90 degrees counter clockwise from the A4C position, the A2C can be obtained by the indicator



(a) (b) (c) (d)

Fig 3. 2 Apical Angles (AA) view including 4 viewpoints corresponding to different apical chamber. (a) the transducer location of AA on the chest; (b) The AA plane of echo beam; (c) four viewpoints cross-section drawn; (d) visual imaging on the screen.

at around 12 to 1 o'clock, as the orientation 2 indicated in Figure 3.2 (b).

The A2C viewpoint is useful to assess the walls motion of the LV, which displays two chambers including left ventricle and left atrium (LA) in this viewpoint, as illustration in Figure 3.3. For LV, anterior wall, inferior wall and apex can be detected. So it is excellent for assessing the motion states of these structures. In addition, Mitral valve (MV) can be viewed between two chambers, and then the motional information can be obtained from this viewpoint.

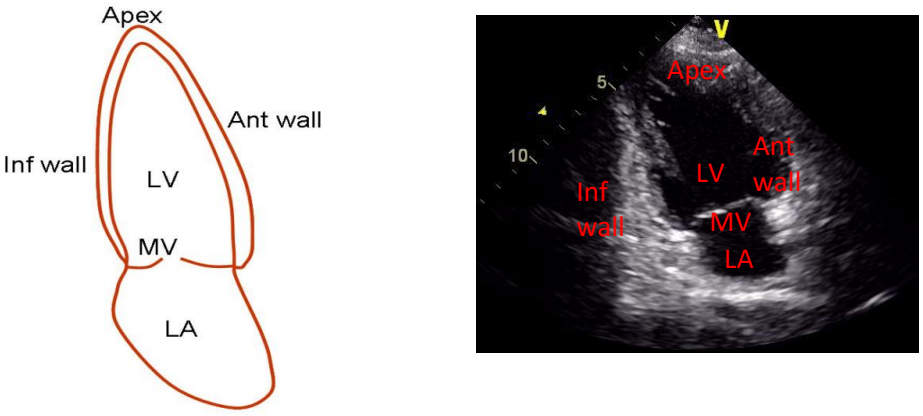


Fig 3. 3 The standard apical 2 chamber viewpoint. (a) Structural map; (b) Normal structures of A2C, i.e. LV and LA.

In actual sampling process, different motion stages can be recorded in the image sequence. In this viewpoint, the LV, LA and MV have different motion stages in two motion phases (the systolic and diastolic processes). The two leaflets of MV are constantly in motion of opening and closing during these two phases of LV. There are some sampled frames from our dataset showing different appearances, stages and image quality of A2C (shown in Figure 3.4).

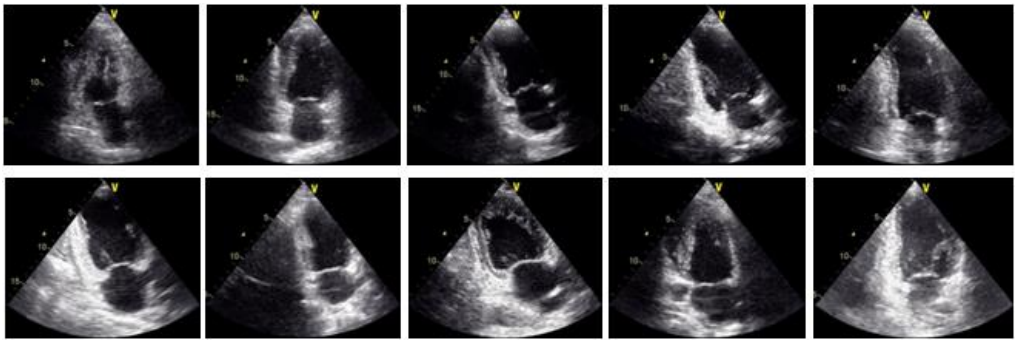
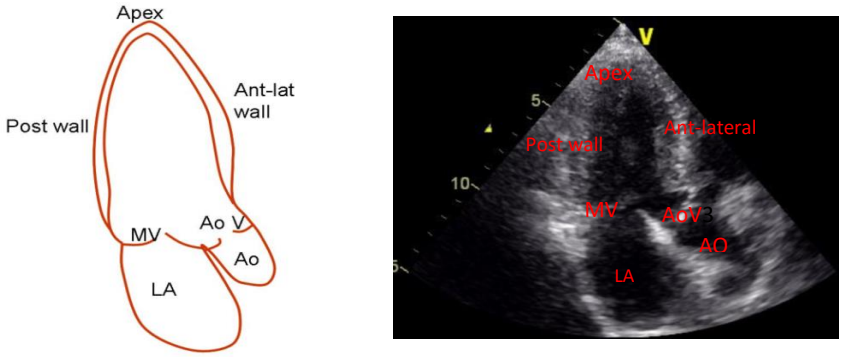


Fig 3. 4 Sample frames from some actual images corresponding to A2C viewpoint. The first two rows are in the ventricular systole with the MV in a closed state. The following two rows are in the ventricular diastole with the MV opening.

The A3C viewpoint is often used to assess the contractility of the anterior-lateral and posterior walls (PW). In comparison with A2C, it has an additional chamber to

the viewpoint, i.e. aorta (AO) and aortic valve (AoV), as shown in Figure 3.5. This view shows similar structures to the parasternal long axis except that the LV apex is well visualized and is in the near field. The LV wall segments and views of the aortic and mitral valves are the same as that in A2C.



(a)

Fig 3. 5 The standard apical 3 chamber viewpoint. (a) Structural map; (b) Actual structures of A3C including LV, LA and AO.

In Figure 3.6, there are sequential images acquired at this viewpoint sampled from the practical detection. In systole of the heart, the LV filled with blood begins to contract, at the same time the MV closes. The AoV then opens and the blood ejects into the AO from the LV. In diastole, a reverse process takes place.

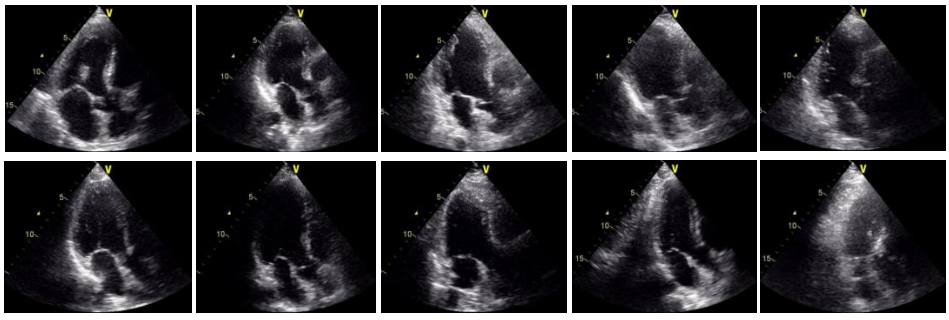


Fig 3. 6 Sample frames from some actual images corresponding to A3C viewpoint. The first two rows are in the ventricular systole with the MV in a closed state and AoV opening. The following two rows are in the ventricular diastole with the MV opening and AoV closing.

In the standard *A4C viewpoint*, all four major chambers of the heart can be seen as shown in Figure 3.7. In addition to the LV lateral wall, apex and septum typically laid out, the right ventricle (RV), the right atrium (RA) and tricuspid valve (TV) are viewed clearly. In the appropriate orientation, the septum (between LV and RV) lines up vertically near the centre of the screen. The LV and LV apex should be vertically oriented and the LV should be approximately parabolic in shape. Sometimes the LV apex appears round or is off the center in only this view, which can possibly affect recognition. This view is also used to assess RV size and function, atrial size, abnormal intra-atrial and inter-ventricular septal movement, as well as diastolic function. A few institutions have standardized this view in right-left reverse with the left ventricle on the left side of the screen and the right ventricle on the right of the screen. The majority of hospitals otherwise use the conventional way we show as follows. The data of this view in the following is aiming at the general case shown in Figure 3.7(b).

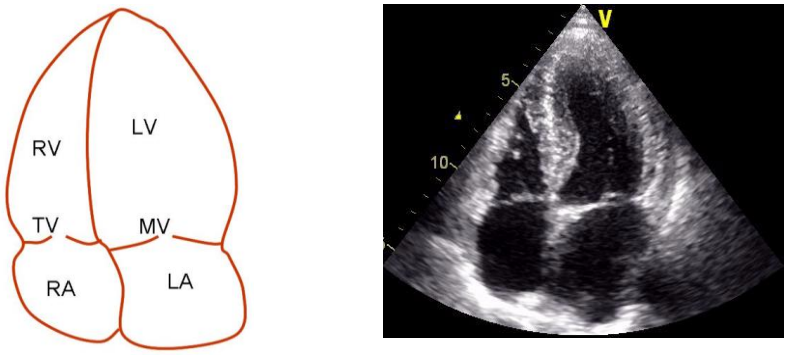


Fig 3. 7 The standard apical 4 chamber viewpoint. (a) Structural map; (b) Actual structures of A4C including LV, LA, RV, RA and septum and valves between these four chambers.

In this viewpoint, ventricle and atrium constitute a kind of motion sequence. The motion of LV and LA synchronizes with that of the RV and RA in a cardiac circle. The MV and TV open simultaneously in cardiac diastole and the blood ejects from

atrium to ventricle. Then it closes during the systole. This synchronous process is indicated in Figure 3.8.

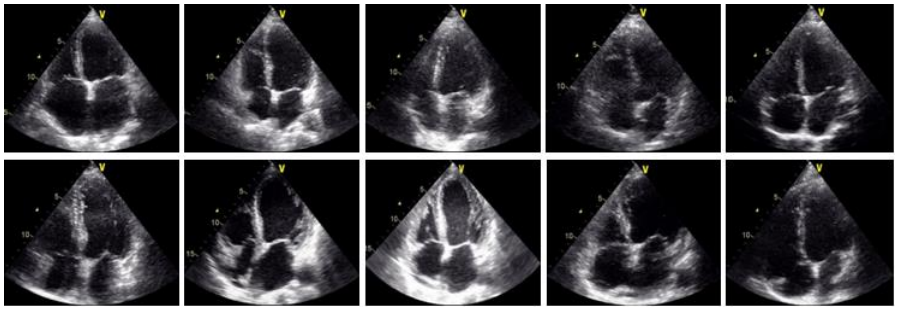


Fig 3. 8 Sample frames from some actual images corresponding to A4C viewpoint. The first two rows are in the systole with the MV and TV in a closed state. The following two rows are in the diastole with the MV and TV opening.

The A5C viewpoint differs from the apical 4-chamber view because of the presence of the aortic valve in the centre of the image (shown in Figure 3.9). It is a good window to assess only the aortic valve and left ventricular outflow tract (LVOT) and their relations to the interventricular septum (IVS) and mitral valve.

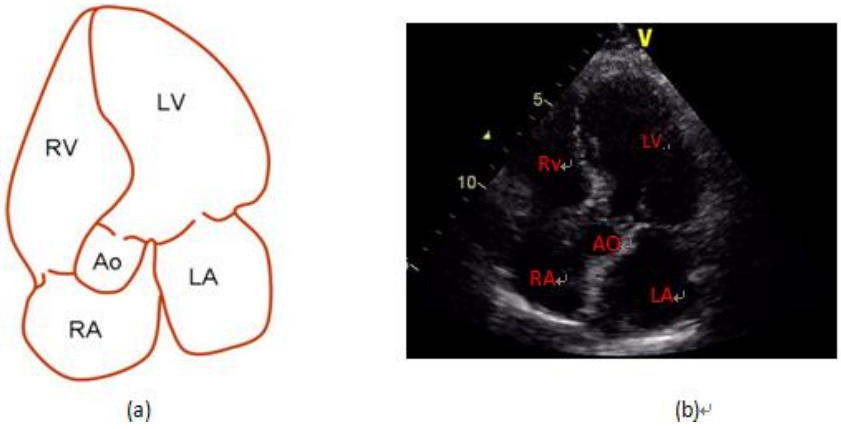


Fig 3. 9 The standard apical 5 chamber viewpoint. (a) Structural map; (b) Actual structures of A5C. On the base of A4C, there is an additional chamber in the center of image, i.e. the aorta.

The five chambers in this viewpoint encompass four real chambers and the AO structure, as indicated in Figure 3.10. In normal process of diastole and systole, the motion status of the four chambers (LV, LA, RV and RA) has been illustrated. The

difference between A5C and A4C is introducing the movement of AO. The AoV will open when the blood ejected from LV into AO in diastole, and then close in the following cardiac systole. In the actual echocardiogram videos, this viewpoint is one of the most fickle ones because of the cardiac translation and rotation. Sometimes the AO becomes blurry or even disappearing. When it happens, A5C video images look like the ones from A4C viewpoint.

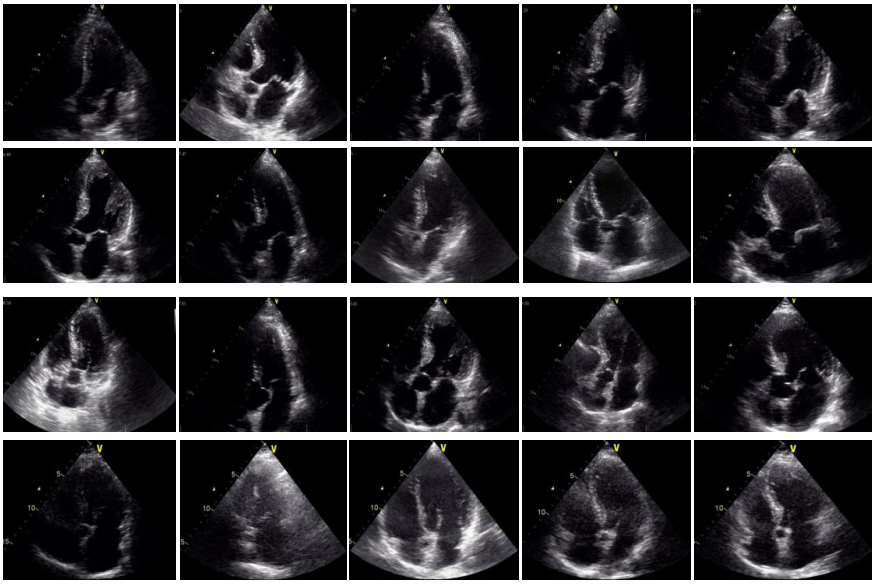


Fig 3. 10 Sample frames from some actual images corresponding to A5C viewpoint. The first two rows are in the systole with the MV and TV closing, while AoV opening or about to open. The following two rows are in the diastole with the MV and TV opening and the AoV closing.

3.2.2 Parasternal long axis view

The parasternal long axis view is generally the first view obtained in a routine transthoracic echocardiogram[115]. In the process of detecting, the probe is located next to the sternum, between the 3rd and 5th intercostal spaces. The notch on the probe should face toward the right shoulder and at about 10 o'clock shown in Figure 3.11 (a). By manipulating the tip of the probe (shown in Figure 3.11 (b) and (c)), PLA standard viewpoint can be obtained, as shown in Figure 3.11 (d).

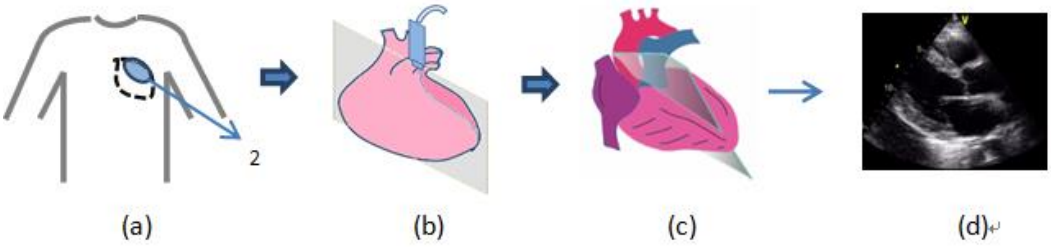


Fig 3. 11 Parasternal long axis view. (a) The transducer location of PLA on the chest; (b) The echo beam slice ripping into the cardiac PLA; (c) The PLA sectorial cross-section drawn; (d) Visual imaging on the screen.

The RV is on the top, close to the probe as indicated in Figure 3.12. The RV dilatation and contractility can be obtained from this viewpoint. The LV should be oriented almost horizontally, where interventricular septum (IVS) and posterior wall (PW) of LV are visualized. In this viewpoint, the leaflets of MV can be displayed clearly as well as aortic valve. So this viewpoint is the best not only for the measurement of the size and walls thickness of the LV but for the observation of the MV motion and the assessment of aortic stenosis. In brief, it is one of the basic viewpoints with significant information in echocardiogram imaging.

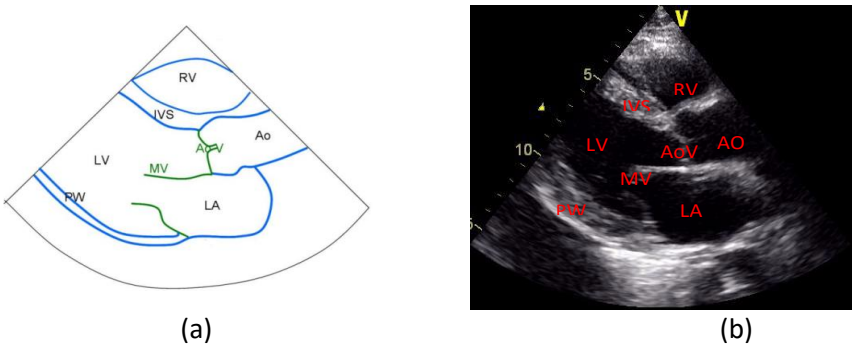


Fig 3. 12 The standard parasternal long axis viewpoint. (a) Simulated structural map; (b) Actual structures of PLA. On the base of A4C, there is an additional chamber in the center of image, i.e. the aorta.

Unlike the apical view images, the images of the PLA viewpoint, characterized by the horizontally oriented LV, are not easily confused with the images corresponding to other viewpoints. The variation about the LV and LA in a cardiac circle is mentioned above. The MV and AoV are in the alternation of opening and closing states with the cardiac contraction and relaxation. Figure 3.13 shows some frames sampled from PLA dataset. Although corresponding to the same viewpoint, all of these images reflect different conditions and stages of the moving heart.

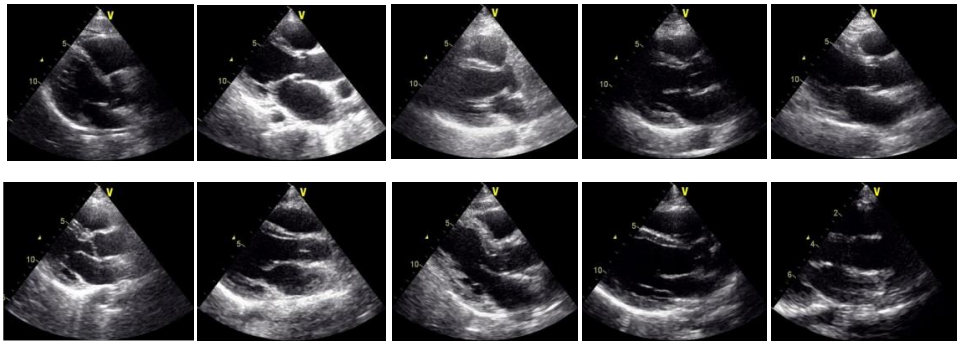


Fig 3. 13 Sample frames from some actual images corresponding to PLA viewpoint. The first row is in the systole with the MV closing, while AoV opening. The following row is in the diastole with the MV opening and the AoV closing.

3.2.3 Parasternal short axis view

On the base of PLA, turn the probe clockwise until the notch on the probe oriented at about 2 o'clock (toward the left shoulder of the body), the parasternal short axis view can be obtained (shown in Fig 3.14 (a) and (b)). From this position, three different viewpoints of the LV can be detected by tilting the probe handle up and down. The most basal (level A in Fig 3.14(c)) viewpoint lays out some valves including the AV, pulmonic valve (PV) and TV. Other standard viewpoints correspond to the LV at the mitral valve level and the mid-ventricle level with papillary muscles and the apex, as indicated in Fig 3.14 (c)-level B and C.

Three viewpoints of PSA view encompass parasternal short axis - aorta (PSAA), parasternal short axis - mitral (PSAM) and parasternal short axis - papillary (PSAP). In fact, PSAA viewpoint corresponds to the basal level, therefore sometimes it is expressed as parasternal short axis - basal (PSAB).

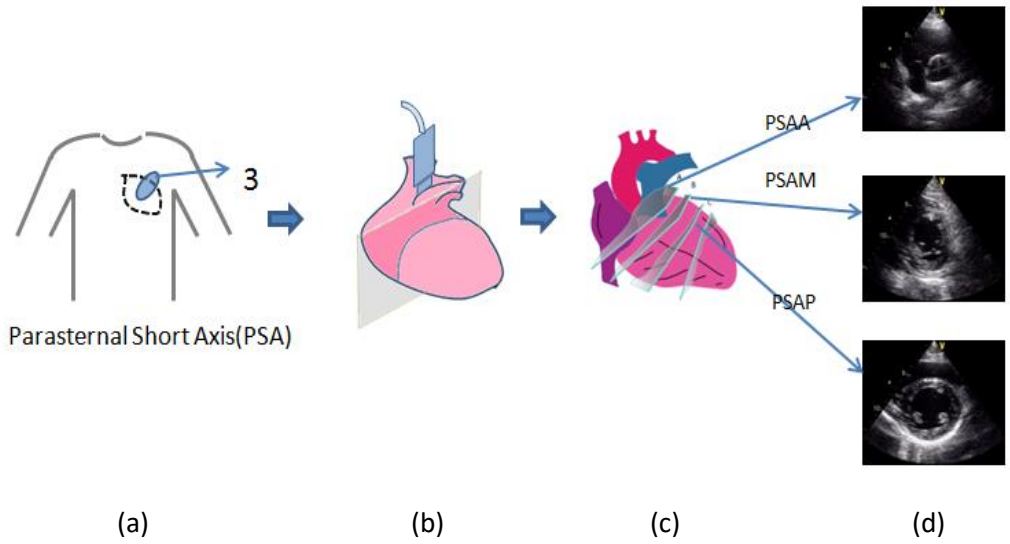
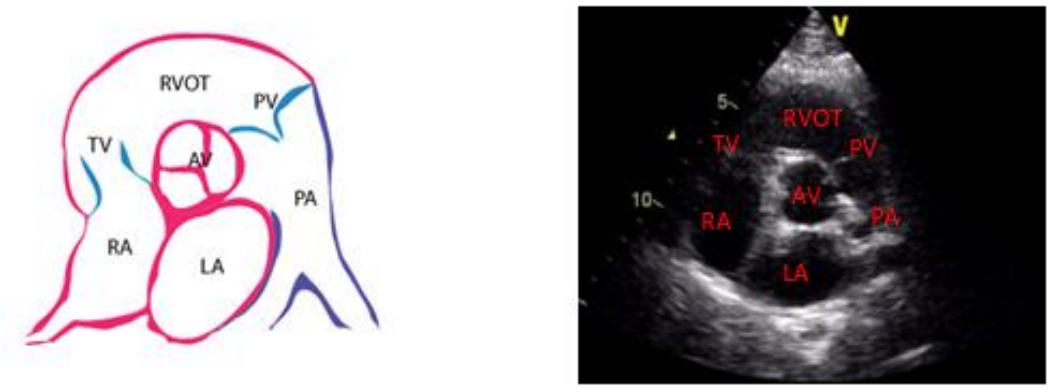


Fig 3. 14 Parasternal short axis view. (a) The transducer location of PSA on the chest; (b) The echo beam slice ripping into the cardiac PSA; (c) The PSA cross-section drawn with three levels (A ~C); (d) Visual imaging on the screen.

The parasternal short axis-aorta (PSAA) viewpoint is the most basal short axis view that lays out the 2 atria (LA and RA), 3 valves (TV, AV and pulmonic valve (PV)) and the RV outflow tract (RVOT). The RVOT is on the top of the image near the probe, and the aortic valve is located at the center of the image with three cusps like 'Y' shape. The PV is on the right of the image (at 1~2 o'clock), connecting the RV to the main pulmonary artery, while TV is located at 9 o'clock on the left of the image, between the RA and RV. LA is located behind the AV, on the bottom of the image. The structure of this viewpoint is indicated in Figure 3.15. It is a good position to look for a pulmonic regurgitation or stenosis and tricuspid regurgitation, as well as to estimate RV systolic pressure [116].

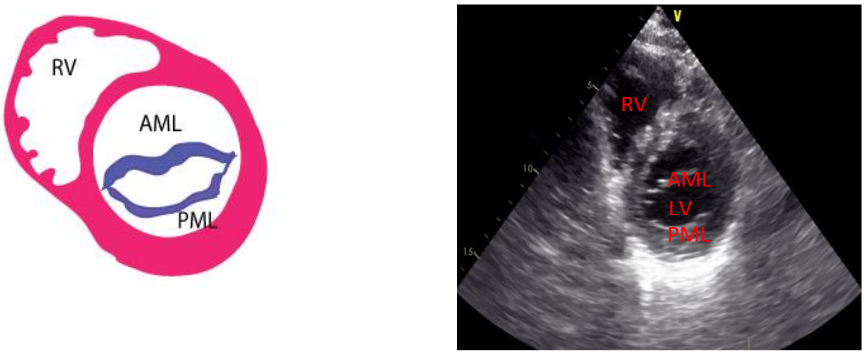


(a) (b)

Fig 3. 15 The standard parasternal short axis-aorta viewpoint. (a) Simulated structural map; (b) Actual structures of PSAA with two aorta and three valves.

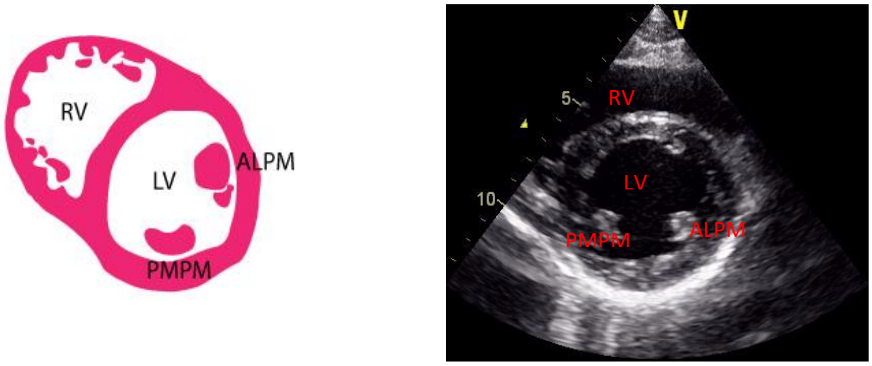
The parasternal short axis-mitral (PSAM) viewpoint can be obtained by tilting the probe tip downward on the base of PSAA viewpoint. Similar to PSAA, the RV is on the top of the image. In the center of the image, it is MV, which looks like a 'fish-mouth' in diastole, the top part is anterior mitral leaflet (AML) which is close to the RV and the bottom one is posterior mitral leaflet (PML), as indicated in Figure 3.16. In the proper state, the LV should be round. But sometimes it is oval. This viewpoint can be used to determine the origin of the regurgitant jet when putting colour Doppler [116].

The parasternal short axis-papillary (PSAP) viewpoint is viewed at mid-ventricle level by tilting the probe tip sequentially away from the basal level, which encompasses a round-shape LV and the papillary muscles (antero-lateral papillary muscle (ALPM) on the right and postero-medical papillary muscle (PMPM) on the left of the image) (shown in Figure 3.17). This viewpoint is important to assess the function of the left ventricle and the contraction of the different walls of the LV.



(a)

Fig 3. 16 The standard parasternal short axis-mitral viewpoint. (a) Simulated structural map; (b) Actual structures of PSAM in diastole.



(a)

Fig 3. 17 The standard parasternal short axis-papillary viewpoint. (a) Simulated structural map; (b) Actual structures of PSAP with two papillary muscles.

The actual sample frames corresponding to three primary viewpoints of PSA view are indicated in Figure 3.18. Although image resolution is not ideal, the PSAA viewpoint has obvious characteristics compared with the other two viewpoints. It can reveal the mutual movement among multiple chambers and valves as shown in Figure 3.18 (a).

While in the PSAM and PSAP viewpoints, the LV and RV are mainly displayed. The major difference between the two viewpoints is the structure shown in the LV. In the PSAM viewpoint, it shows the motion stage of the MV with opening and closing as the 'fish-mouth' appearance. For the PSAP viewpoint, papillary muscles

can be displayed, which attach to the left ventricular wall with the movement of contraction and relaxation. Figure 3.18 (b) and (c) illustrate some real sample images about the PSAM and PSAP, sometimes they can be confused easily because of the influence of some factors, such as the cardiac translation and rotation, the speckle noise and sonographer experience.

(a) Sample frames from the PSAA viewpoint

(b) Sample frames from the PSAM viewpoint

(c) Sample frames from the PSAP viewpoint

Fig 3. 18 Actual sample frames from PSA view. (a) Sample frames from different videos of the PSAA viewpoint; (b) Sample frames from the PSAM viewpoint. The first two images are in cardiac systole, others correspond to diastole with the MV as 'fish-mouth' shape; (c) Sample frames from the PSAP viewpoint. The first two images are in cardiac systole, others correspond to cardiac diastole.

4. Methodology

4.1 Overview of the work carried out in this study

This chapter introduces a novel spatial-temporal method for local features in echocardiogram video. The fundamental purpose is to generate local features that exhibit high repeatability and distinctiveness under different echocardiogram viewpoint variations. In view of the good performance of 2D KAZE feature in classifying multi-class viewpoints of echocardiogram video [68], we extend this technology into 3D (2D spatial and 1D temporal) and generate 3D KAZE feature to represent video sequences. In this study, the 3D KAZE feature detection is implemented as shown in Figure 4.1.

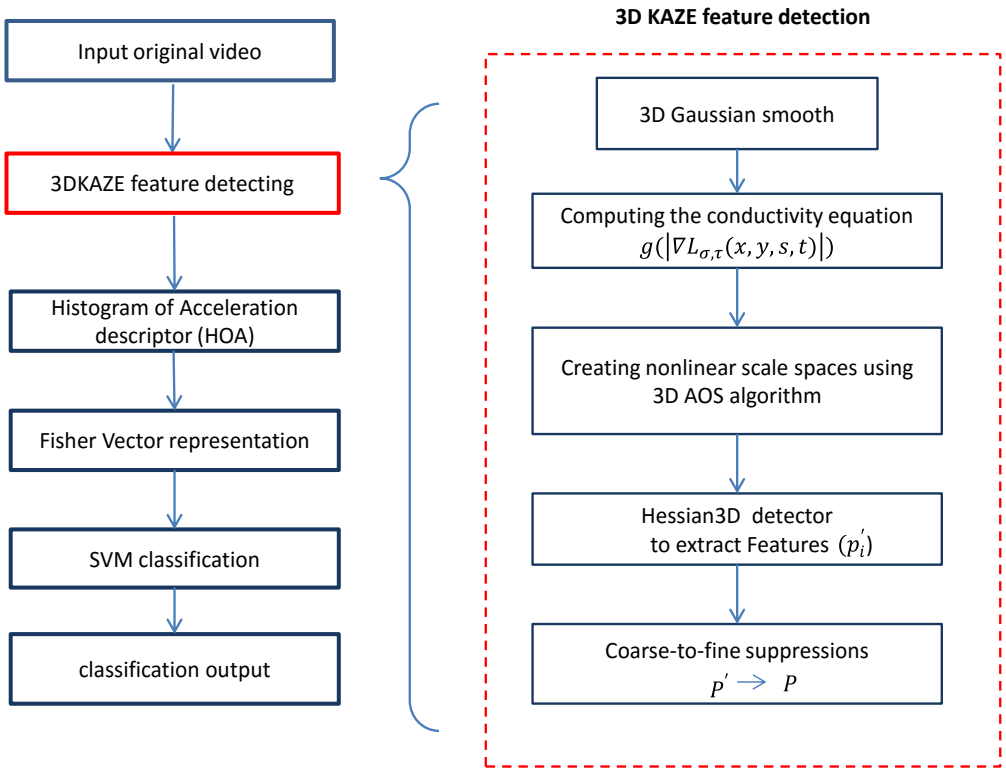


Fig 4. 1 The flowchart of echocardiogram video classification in this study. The red box corresponds to 3D KAZE feature detecting process.

4.2 The spatial-temporal KAZE feature detection

4.2.1 Echocardiogram video pre-processing

The ultrasound videos that we are concerned with come from the original video data exported from clinical detection, which have three colour channels (RGB) corresponding to the hue, saturation and luminance respectively. The area of each video displaying the actual cardiac structural and motional information is referred to as the fan area due to its shape and the scanning principle of the ultrasound transducer. The real size of this fan area in a given ultrasound video depends on the ultrasound machine and its settings. In our work, all the videos have the same resolution. Since the artefacts including the patient's personal information, diagnostic date and the name of corresponding hospital, can be indicated in the video sequence. It is no use for the subsequent local feature detection. The fan region containing cardiac structures alone is selected and cropped before the following processes. The pre-processing flow is illustrated in Figure 4.2.

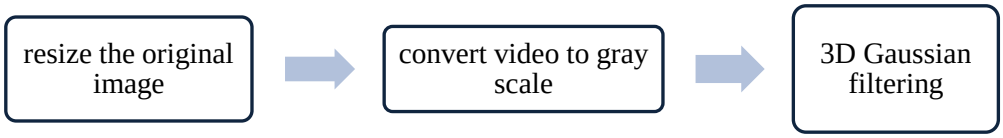


Fig 4. 2 Pre-processing flow of original echocardiogram video

In addition, another important pre-processing stage in echo video detection and analysis is removing noise while revealing cardiac structural details clearly. Gaussian filter, which is a group of low-pass filters passing over low frequency components and reducing high-frequency components [117], is one of the most popular filtering technologies to denoise image. In this research, the echocardiogram video sequence can be viewed as a spatial-temporal volume v to

have variations with various degrees along both spatial and temporal directions. So the spatial-temporal separable Gaussian function [75] is applied to denoise the volume. The filtered volume L is generated by its convolution with the anisotropic Gaussian kernel with independent spatial and temporal variance (σ^2, τ^2):

$$L(x, y, z\ ; \sigma^2, \tau^2) = G(\ x, y, z\ ; \sigma^2, \tau^2) * v(\ x, y, z\ ) \qquad (4.1)$$

where the spatial-temporal separable Gaussian kernel G is defined as:

$$G(x, y, z; \sigma^2, \tau^2) = \frac{1}{\sqrt{(2\pi)^3 \sigma^4 \tau^2}} \exp(-\frac{(x^2+y^2)}{2\sigma^2} - \frac{z^2}{2\tau^2}) \qquad (4.2)$$

Here, the reason we use a separate scale parameter for the spatial and temporal domain blurring is that the spatial and temporal resolution of echocardiogram video sequence are generally independent. In our experiment, on the base of the separability of Gaussian function, the spatial-temporal Gaussian filtering process is completed from three independent dimensional calculations including two spatial directions (horizontal and vertical directions) and one temporal direction. The variances corresponding to spatial and temporal directions are set to be constant（σ_0， τ_0）.

4.2.2 Computing the conductivity equation of nonlinear diffusion filtering

The Echocardiogram video sequences in our research are space-time frames with 2-dimentional spatial and 1-dimentional temporal information. It can be viewed as a generalized 3D volume by prolonging the 2D spatial domain along the temporal axis. So we can apply multiscale analysis for spatial-temporal sequences the same way as for a 3D volume. As defined in[118], the image multiscale analysis is a family of transforms which change original images (original 3D volume) into a number of simplified images depending on a set of scale parameters. Nonlinear

scale space as a representation of multiscale analysis can make filtering locally adaptive to the image data. Its realization depends on nonlinear diffusion.

Nonlinear diffusion approaches describe the evolution of the luminance of an image through increasing scale levels as the divergence of a certain *flow* function that controls the diffusion process. This process is realized by using nonlinear partial differential equations (PDEs). One of the first attempts to derive a model within a PDEs framework is conducted by Perona and Malik [119]. They propose a nonlinear anisotropic diffusion model so as to remove noises and highlight the edge regions based on the traditional heat conduction equation:

$$\begin{cases} \frac{\partial u}{\partial t} = div(C(x, y, z\,, t) \cdot \nabla u) \\ u|_{t=0} = u_0 \end{cases} \tag{4.3}$$

where u_0 is the original image, and div and ∇ are the divergence and gradient operator respectively. The diffusion coefficient C can make the filtering adaptive to the local image structure, which depends on the choice of scale parameter t. Increasing t leads to simpler image representations.

In our research, given an echocardiogram 3D volume $v\colon R^2 \times R \to R$, the diffusion coefficient is chosen to be an appropriate function of the estimate of the gradient [120] :

$$C(x, y, z, t) = g\left(\left\|\nabla G_{\sigma,\tau} * v(x, y, z)\right\|\right) \tag{4.4}$$

Where $G_{\sigma,\tau}$ is the spatial-temporal separable Gaussian kernel, which is defined as Eq (4.2). According to the property of convolution in Eq. (4.5):

$$\nabla G_{\sigma,\tau} * v(x, y, z) = G_{\sigma,\tau} * \nabla\, v(x, y, z) \tag{4.5}$$

The term in the right side represents the smoothing of the spatial-temporal gradient volume. g is the coefficient function, which has two widely used properties proposed in [119]:

$$g_1(\|\nabla(x,y,z)\|) = \exp\left(-\left(\frac{\|\nabla(x,y,z)\|}{K}\right)^2\right) \tag{4.6}$$

$$g_2(\|\nabla(x,y,z)\|) = \frac{1}{1+\left(\frac{\|\nabla(x,y,z)\|}{K}\right)^2} \tag{4.7}$$

Where K is the contrast parameter to control the smooth level. For increasing value of K, only higher gradients are considered, and the detail will be given in the next section. These two models making the diffusivity has to be such that $g(\|\nabla(x,y,z)\|) \to 0$ when $\|\nabla(x,y,z)\| \to \infty$ and $g(\|\nabla(x,y,z)\|) \to 1$ when $\|\nabla(x,y,z)\| \to 0$. Since the term $\left\|\nabla G_{\sigma,\tau} * v(x,y,z)\right\|$ meets the conditions, the diffusion coefficient (equation 4.4) makes the diffusion process more stable with respect to noise ([121], [122]). When setting the coefficient function g to be equal to 1, our framework can be viewed as the Gaussian scale space (linear scale space). For most of the voxels, it works like the Gaussian diffusion to remove the noises away. While for the strong edges that correspond to the cardiac structural boundaries, our framework can highlight the boundaries.

4.2.3 Setting the contrast parameter and evolution times of 3D KAZE

4.2.3.1 Setting the contrast parameter

For the echocardiogram image, different viewpoints can present various cardiac structures such as cavities and valves. In ultrasound imaging acquisition, these cavities are indicated corresponding to lower intensity area, while all valves and

atrial septum are grouped between cavities with high intensity. The diastolic and systolic motion states corresponding different cardiac structures can be shown with the change of intensity among the regions. This kind of region is primarily corresponding to the adjacent boundary regions of cardiac structures, which can reflect a large amount of structural and motional information of the heart. So in order to preserve details in these areas as much as possible during smoothing process, the contrast of the image should be considered in choosing the conductivity function of filtering. The contrast parameter K is introduced in [119], which is the difference in the image intensities on the left and right of the boundary, determines which edges have to be enhanced and which ones have to be cancelled. It can be fixed by hand or automatically by means of image gradient.

Alcantarilla et al. [67] obtained the gradient histogram from a smoothed version of the original image and took the 70 percentile of its intensity level as the contrast parameter. It is an empirical result and cannot be applied to all kinds of images. As is well known, standard deviation σ and variance σ^2 can be applied to measure the dispersion in image distribution, which are more preferable as contrast determinant[123]. In order to generate the contrast parameter automatically, we aim to not only capture the dynamic range of grey level but also reflect the distribution of lower and higher intensity area in global fashion of whole echocardiogram videos. In this research, the variance of grey levels σ^2 is implemented to reflect the distribution of images in echocardiogram video sequence as formulated in Eq. (4.8).

$$\sigma^2 = \sum_{i=1}^{M} P_i (I_i - \bar{I})^2 \tag{4.8}$$

Where I_i and $\bar{I}$ are the actual grey value and the corresponding mean in echocardiogram video, and the probability of the i_{th} grey value can be calculated as P_i. M stands for the size of grey levels.

4.2.3.2 Evolution time

The transformation in linear scale space (such as Gaussian smooth) can be viewed as a process of diffusion filtering for a certain time. It is illustrated in some researches ([124], [67]) that Gaussian convolution (linear scale space) with standard deviation σ_0 (in pixels) is equivalent to filtering the image for some time $t_0 = \sigma_0^2/2$ (called evolution time). In the process of building nonlinear space, we apply this conversion to generate a set of evolution times, and consequently obtain scale parameters of nonlinear scale space. The transformation relationship can be shown through the following mapping:

$$t_i = \frac{1}{2}\sigma_i^2, \quad i = \{0 \dots \dots N\} \tag{4.9}$$

Where N is the scale level. $\sigma_i = \sigma_0 2^{\frac{i}{4}}$, σ_0 is the base level of the spatial scales (with 1.6 in our experiment). For temporal space, it has the same relationship, i.e. $t_i = {\tau_i}^2/2$, $\tau_i = \tau_0 2^{\frac{i}{4}}$ (τ_0 is the base level of the temporal scales with 1.6 in our experiment).

4.2.4 Creation of nonlinear scale spaces

In what way that the nonlinear scale diffusion Eq. 4.3 can be solved accurately is the next step to consider. Although there have been a number of proposals introducing numerical schemes for anisotropic diffusion processes, e.g. in [125], [126], [127], [128], probably there are two most popular ways to implement anisotropic diffusion filters which are explicit and semi-implicit schemes [129]. The former is simple, straight-forward, and computationally cheap because only matrix-vector multiplications are required without solving linear or nonlinear system of equations. However, it is conditionally stable and limited to small time steps as stated by Barash et al.[130], Grewening et al. [129], Weickert et al.[131]. They

present the comparison and analysis about these two schemes in details and propose that semi-implicit Additive Operator Splitting (AOS) scheme can conditionally satisfy all the requirements of discrete scale-space and remain absolutely stable.

The discretization of equation 4.3 can be expressed as:

$$\frac{v^{i+1}-v^{i}}{\tau} = \sum_{l=1}^{m} C_l(v^i)v^{i+1} \tag{4.10}$$

Where C_l is a matrix that encodes the image conductivities for each dimension, derived from Eq.(4.4). m is the size of image dimensions (default 3 for video). τ is the time step size. The solution v^{i+1}can be expressed as:

$$v^{i+1} = \frac{1}{m}\sum_{l=1}^{m}(I - m\tau C_l(v^i))^{-1}v^i \tag{4.11}$$

In this semi-implicit scheme, it is necessary to solve a linear system of equations, where the system matrix is tridiagonal and diagonally dominant. Such systems can be solved very easily and efficiently by means of the well-known Gaussian elimination algorithm (so-called Thomas algorithm), as applied in [67].

4.3 Feature detection with Hessian saliency measures

For the Hessian saliency detection, the determinant of 3D Hessian matrix is used to measure the saliency of the spatial-temporal volume. The matrix is shown as:

$$H(\cdot;\sigma,\tau) = \begin{pmatrix} L_{xx} & L_{xy} & L_{xt} \\ L_{yx} & L_{yy} & L_{yt} \\ L_{tx} & L_{ty} & L_{tt} \end{pmatrix} \tag{4.12}$$

The strength of each feature point at a certain scale is given by the determinant of its Hessian matrix $\det(H)$. By using the scale-normalized spatial-temporal Laplacian [73] localize final feature points in the scale space as local maxima of

$$S = \sigma^{2p}\tau^{2q}\det(H) \tag{4.13}$$

Where (L_{xx}, L_{yy}, L_{tt}) are the second-order derivatives of the video sequence v in the Gaussian scale space. In our research, we calculate it in each nonlinear scale space obtained in Section 4.2, and view these maxima as the candidate feature point p'. Parameters of (p, q) are constants and fixed as (2, 1) in our experiment. (σ, τ) are independent spatial and temporal values in nonlinear scale space as mentioned above. The determinant of the 3D Hessian matrix is computed on each of the spatial and temporal scale respectively, which is the first step to detect features in position space.

In order to ensure the representativeness and stability of the features, we consider that a real feature should have strong response between neighbouring scale spaces as well. Therefore, we introduce a scale-normalized spatial-temporal Laplacian operator [74] to search the saliency in two neighbouring scale spaces based on all candidate points. The operator F is formulated by

$$F\left(p^{'}, \sigma_i, \tau_i\right) = \sigma_i^2 \tau_i^{1/2}(L_{xx} + L_{yy}) + \sigma_i \tau_i^{3/2} L_{tt} \qquad (4.14)$$

$$F\left(p^{'}, \sigma_i, \tau_i\right) > F\left(p^{'}, \sigma_l, \tau_l\right) \qquad l \in \{i-1, i+1\} \qquad (4.15)$$

The candidate feature point p', whose Laplacian response $F(p', \sigma_i, \tau_i)$ meets the condition as Eq. (4.15), can be viewed as a feature point, and the corresponding scale is called characteristic scale.

In addition, for further discarding unreal features caused by noises in echocardiogram video sequence, we propose a coarse-to-fine strategy to discard the unreal feature points in scale spaces and spatial location. In each characteristic scale space, since the gradient magnitude can reflect the boundary information of the cardiac structures. We introduce the gradient magnitude into detecting process. The mean value of the gradient magnitude of the echocardiogram video sequence is

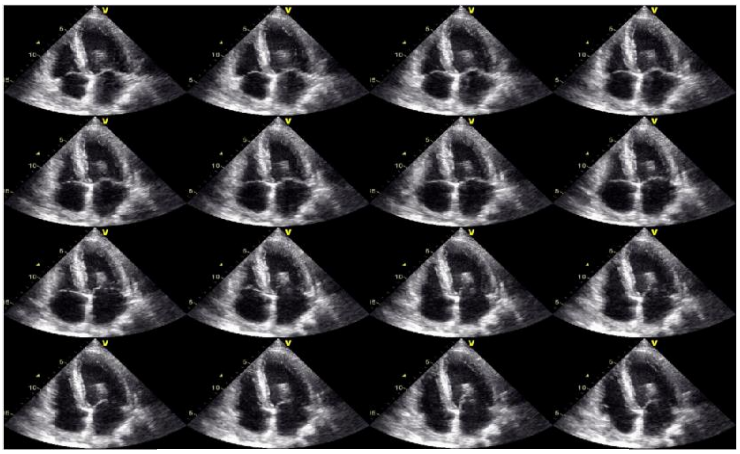
calculated and viewed as the threshold to discard those points whose gradient magnitude is lower than the threshold. This step is called the lower magnitude suppression in our research.

In reality, in real image domain, some spatial-temporal feature points detected as mentioned above are concentrated on a small region with the same characteristics. Therefore, it is little use for representation and classification these regions rather than time-consuming in detection. We search the neighborhood of each feature points and discard those points whose grey level is lower than the variance of original video sequence. The variance is calculated as expressed in Eq. (4.8). This process is called neighborhood suppression. The corresponding results are illustrated in the following chapter.

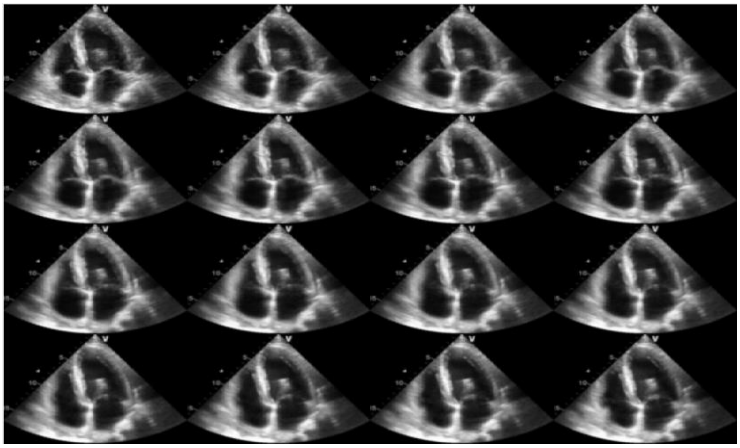
5. The results of 3D KAZE feature detection

5.1 Pre-processing result

The region of interest (ROI) for every echocardiogram videos in this research is cropped empirically with the size of 341×415. Considering the time-consuming factor, the hue and saturation information in echocardiogram video classification are not necessary for cardiac structure detection and are eliminated. Instead, only intensity information is applied for the subsequent calculations. The result of filtering for echocardiogram video sequence is indicated in Figure 5.1.



Original consecutive video sequence



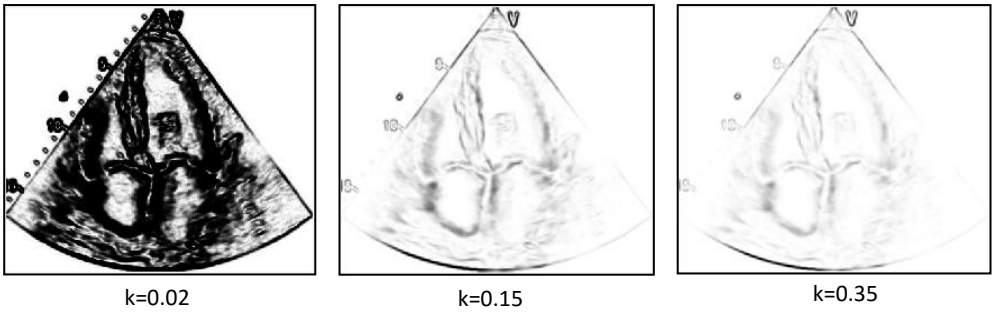
Gaussian filtered image volume

Fig 5. 1 The result of applying spatial-temporal Gaussian filter. The top diagram shows 16 consecutive frames in one of the original echocardiogram videos. The bottom one is the corresponding filtered volume with the variance of (1.6, 1.6).

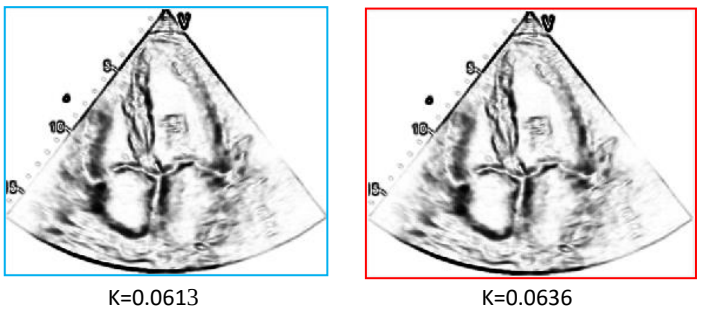
The variances $(1.6, 1.6)$ corresponding to spatial and temporal direction show good filtering effect in our experiment.

5.2 The conductivity result for echocardiogram video

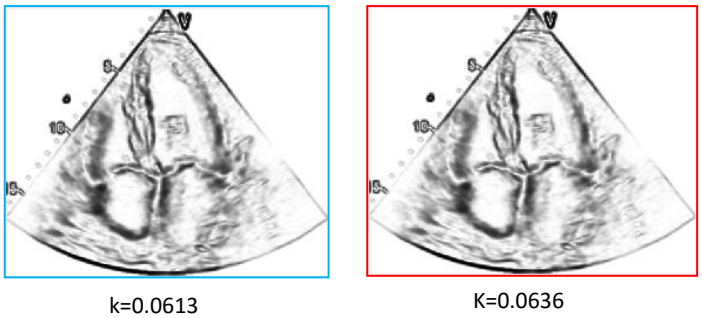
The experimental result confirms that lower gradient areas in echocardiogram video are blurred, while higher gradient regions (e.g. the boundary area) are reserved when continuously increasing contrast parameter K in computing conductivity equation. This variation trend is indicated in Figure 5.2 (a).



(a) Coefficient function g1 with different manual settings of contrast parameter



(b) Coefficient function g1 with histogram(left) and standard deviation(right) methods to calculate the contrast



(c) Coefficient function g2 with histogram (left) and standard deviation(right) methods to calculate the contrast

Fig 5. 2 The conductivity results derived from different contrast parameter K.

We try to use three methods to determine the value of K : manual settings, gradient histogram statistics and the gray level distribution regularities (the variance) respectively. The manual setting method (as shown in Figure 5.2(a)) can provide random values without considering the grey level or gradient of the image (e.g. $k = 0.02, 0.15$ ...), it is fixed and cannot make adaptive adjustment with the different videos. Gradient histogram method (the left with the blue boxes in (b) and (c)) can reflect the magnitude counting of gradient image, it varies with different videos and has some changes when taking different percentiles (e.g. from 70% to 90% in [67]). The variance method in our research (the right side with the red boxes in (b) and (c)) contains the probability statistics and distribution regularities of the grey level in echocardiogram videos. It can be calculated directly from the original video without any parameter settings.

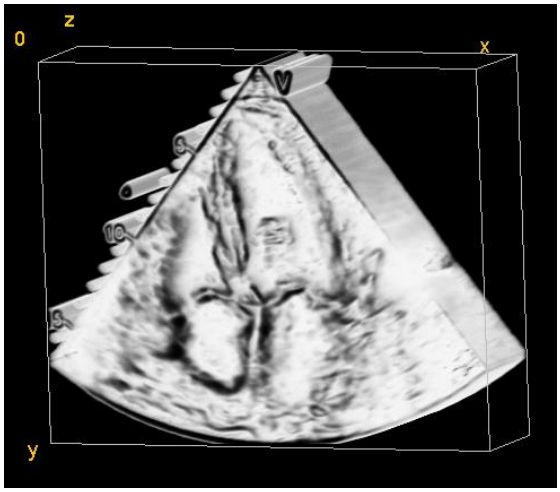


Fig 5. 3 The spatial-temporal conductivity result using coefficient function g_1 with k=0.0636

In our experiment, the coefficient function g_1 (shown in Eq. (4.6)) is applied to blur the image. Compared with the results in Figure 5.2 (b) and (c), the function g_1 promotes high-contrast edges better. From visual perspective, the contrast of the boundary between cardiac structures shown in Figure 5.2 (b) is greater than the

result in Figure. 5.2 (c). It is conducive to highlight the edge information in the further detection. For echocardiogram video sequence, the result of spatial-temporal conductivity using coefficient function g_1 is shown in Figure 5.3.

5.3 The comparison between linear and nonlinear filtering methods

The nonlinear scale spaces corresponding to evolution time (Eq. (4.9)) is obtained based on Eq. (4.11). In Figure 5.4, the comparison result obtained by using Gaussian and nonlinear diffusion methods is exhibited. With the increasing of the scale, Gaussian filtering makes image blurring significantly including noises and structural details, whereas nonlinear filtering remain more structural information.

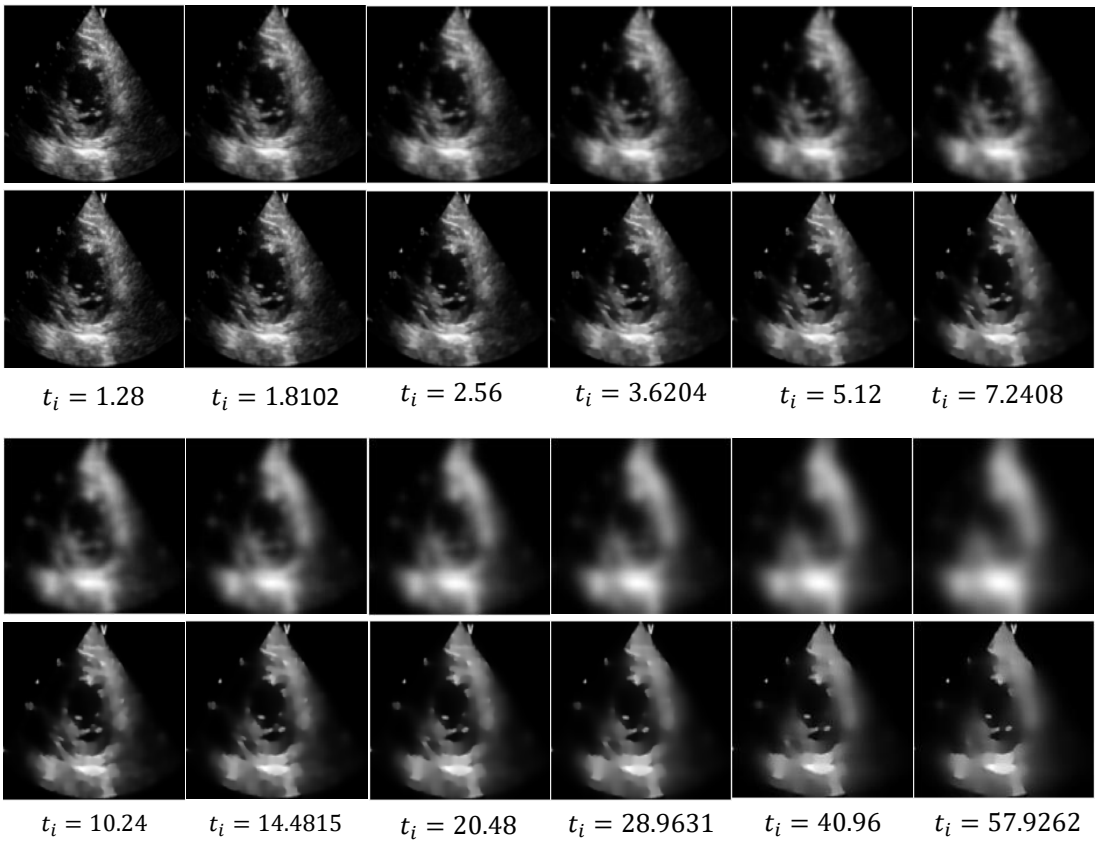


Fig 5. 4 The filtering comparison between the Gaussian and nonlinear diffusion scale spaces. The first and third rows correspond to Gaussian scale space with increasing standard deviation σ_i. The second and fourth rows show nonlinear diffusion scale space with the evolution time t_i. The relationship between σ_i and t_i is stated in Eq. (4.9).

5.4 The result of 3D KAZE feature detection

5.4.1 The detecting algorithm

In order to illustrate the process of 3D KAZE feature detection, we provide an algorithmic description in Figure 5.5. In the actual implementation, the initial values corresponding to the spatial and temporal scales are $\sigma_0 = 1.6$, $\tau_0 = 1.6$, which shows better results as demonstrated in Figure 5.1. The scale level is $N = 12$ in our following experiment settings, which can be changed accordingly.

Input: original video V_0 , σ_0 , τ_0 , evolution time t_i $(i = 1 \dots N)$
Output: feature points P
Build scale space $V = \{V_1, V_2, \dots V_N\}$
Compute the contrast parameter K
For $i = 1$ to N do

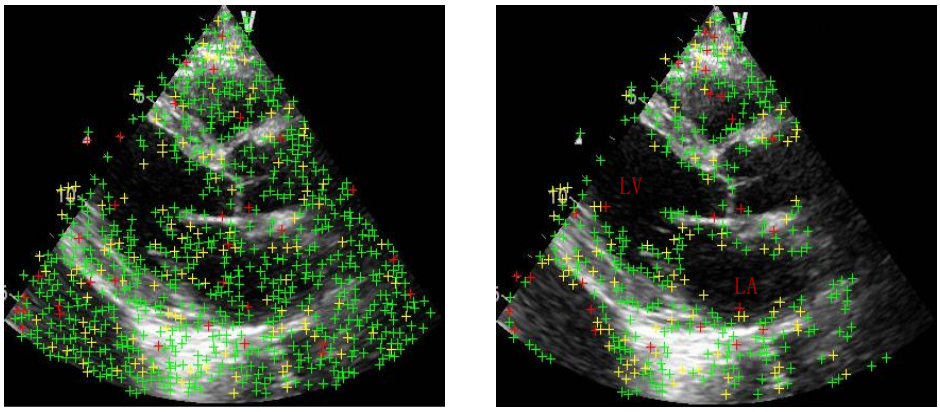
1. Compute conductivity equation C
    - Gaussian convolution: V_{i-1} to G_i
    - Gradient: ∇G_i
    - Magnitude calculation: $\|\nabla G_i\|$
    - Compute conductivity equation C using Eq. (4.4)
2. Compute nonlinear scale space V_i using Eq. (4.11)
3. Compute Hessian matrix H_i
4. Search the local maximum determinant P_i' of H_i

End
Weak feature points' suppression $P' \rightarrow P$

Fig 5. 5 The algorithm of the spatial-temporal KAZE feature points generation.

In addition, we propose lower magnitude and neighborhood suppression strategies to discard less interesting features over all scale spaces. As mentioned above, using Hessian saliency strategy, we can detect a set of candidate feature

points, as shown in Figure 5.6 (a). After suppression processing, many unneeded feature points are discarded especially in local non-boundary regions, e.g. in the LV and LA areas (as shown in Figure 5.6 (b)).



(a) candidate feature points (b) features points after suppression

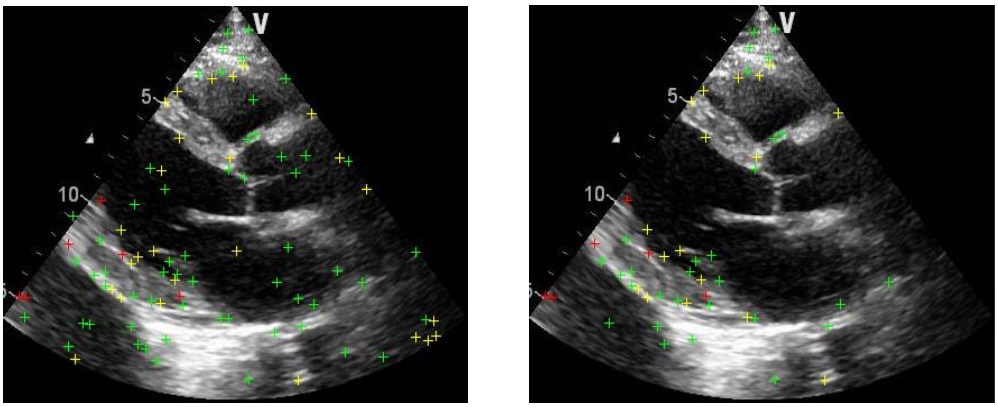
Fig 5. 6 Illustration of spatial-temporal feature points using the Hessian saliency measure and suppression method in multi-scale space. Different color points correspond to different characteristic scales.

5.4.2 Alternative spatial-temporal feature detecting methods

In order to compare with other spatial-temporal feature detection method, we introduce two well-known spatial-temporal detecting strategies to search features in the echocardiogram video sequences in our experiments.

5.4.2.1 Harris3D feature detector

The detection results are illustrated in Figure 5.7 (a). The feature points detected in different scales are indicated by using different colours. In order to discard the weak points that are far away from the boundaries, we apply the lower magnitude suppression method in our experiment, as shown in Figure 5.7 (b).

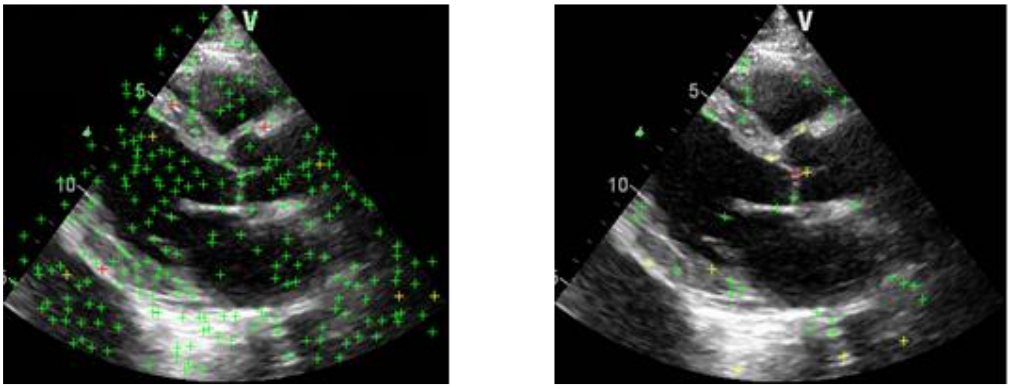


(a) Harris3D feature points (b) Harris3D feature points after

Fig 5. 7 Illustration of spatial-temporal feature points using the Harris3D feature detecting method in multi-scale space (shown in different color).

5.4.2.2 Gabor feature detector

This method originally works in a single scale. In order to compare with other detection method, we extend to multiple scales as mentioned above. The results for detecting an echocardiogram video sequence are indicated in Figure 5.8 (a). The green points correspond to the beginning levels(σ_0, τ_0) , while the yellow and red points are detected in other following scales. Multi-scale detecting has little obvious impact for feature detection in Gabor method. We apply the lower magnitude suppression method to reduce the points beyond the boundaries as shown in Figure 5.8 (b).



(a) Gabor feature points (b) Gabor feature points after suppression

Fig 5. 8 Illustration of spatial-temporal feature points using the Gabor filters in multi-scale space (shown in different color).

5.4.3 The comparison results of detecting strategies

The final feature points for the echocardiogram video sequence are illustrated in Figure 5.9 by using different detecting strategies (corresponding to different rows). All of these results are obtained based on multiple scale spaces. The colour of feature points refers to different scales.

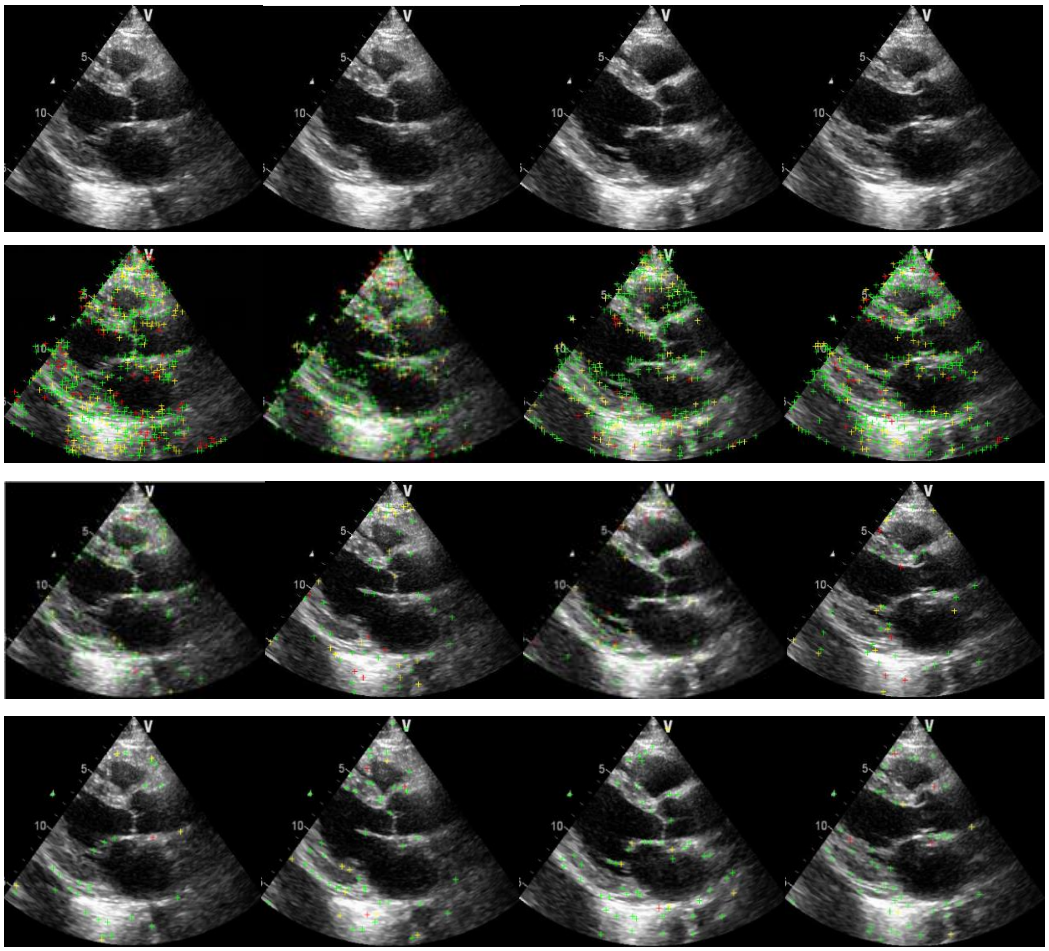


Fig 5. 9 The comparison of feature points detected by different methods: the determinant of Hessian matrix (second row), Harris 3D (third row), Sparse Gabor (fourth row).

The Hessian saliency measure aims at a rather dense feature point detector [73]. So the feature points are located densely at the cardiac structures (e.g. chamber walls and valves shown in the second row of Figure 5.9). For Harris3D detector, it is based on spatial-temporal extension of Harris corner criterion. True spatial-temporal corner points are relatively rare in our echocardiogram video sequence

because of its inherent criterion (without considering boundary detection). Gabor method can detect the features densely (as shown in Figure 5.8 (a)). After suppression process, the majority of feature points are discarded. Sparse features that are too sparse can prove troubling in a recognition framework, as observed by Lowe [63]. For echocardiogram video with four cardiac chambers and some valves, all structural details should be reflected through the feature points. So, we apply the determinant of Hessian matrix to detect local maxima in multiple scale spaces of echocardiogram video sequence in our research.

Table 5.1 summarizes the comparison results between our proposed method and other two detectors. In our experiment, the mean number of generated feature points per frame is illustrated. Our method generates nearly 3 times more features than Harris3D and sparse Gabor detector after suppression. Although dense Gabor3D method can detect denser features than our method, all features are located in the whole region widely and uniformly as shown in Figure 5.8(a) without highlighting the structural characters of the heart.

Detecting method	scale selection	feature set	suppression	mean number per frame
Harris3D	Yes	Sparse	Yes	51
Sparse Gabor3D	Yes	Sparse	Yes	40
Dense Gabor3D	No	Dense	No	194
proposed method	Yes	Dense	Yes	200

Table 5. 1 The comparison using different detecting measures.

5.4.4 The measure of 3D KAZE feature stability

In order to evaluate the detecting method, we measure the stability, i.e. how many spatial-temporal feature points can be detected in both original video sequences and corresponding various geometric and photometric transformations. It can be formulated by

$$S = \frac{\sum(P_{ori} \cap P_{tra})}{\sum P_{ori}} \times 100\% \quad (5.1)$$

Where $P_{ori}$and P_{tra} correspond to the feature points detected in the original video and its transformation. We compare our detecting method with Harris at multiple scales whereas Gabor3D detectors are extracted at both single and multiple scales. The evaluation dataset includes a set of randomly selected echocardiogram video sequences with artificial transformations such as rotation, noisy, blurring and intensity variance. All parameters about the Harris3D detector are the same with our proposed method including suppression process. For Gabor3D method, we divided into two parts: one is the original detecting method applied in [78](is called dense Gabor detector), another is multi-scale detecting with lower magnitude and neighborhood suppression as mentioned in our proposed method, which is known as sparse Gabor detector in our experiment.

Effect of Rotation

We sample randomly some video sequences belonging to different viewpoints and rotate them along the clockwise up to 90 degree (e.g. $5^{\circ}, 10^{\circ}, 20^{\circ}, 40^{\circ}, 60^{\circ}, 90^{\circ}$). Comparing the detecting features in before and after rotational video sequences, an average score is computed using Eq. (5.1).

The effect of rotation is shown in Figure 5.10. All the detecting methods show excellent tolerance to rotation. Dense Gabor feature shows better performance than others. Because it filters video from spatial and temporal space and search the

maxima in the filtered images directly. Rotation transformation has relatively little performance impact on the filtering process. While for Harris and our method, the first or second order derivatives should be calculated in detecting process, which leads to the results that can be influenced more or less with the rotation.

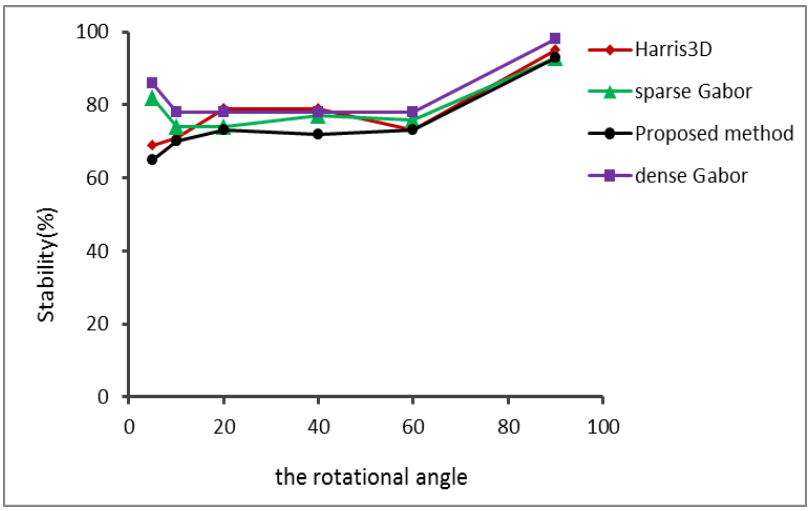
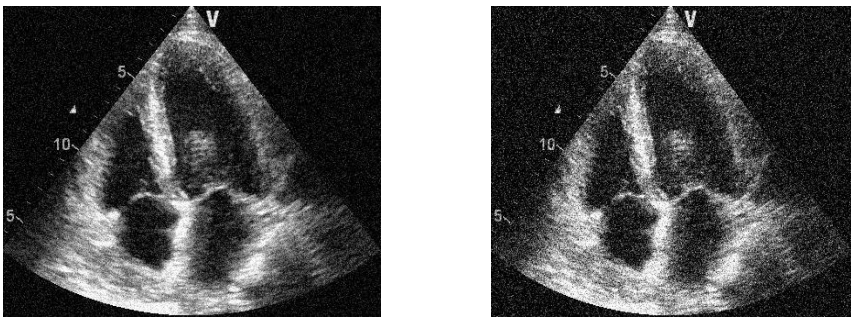


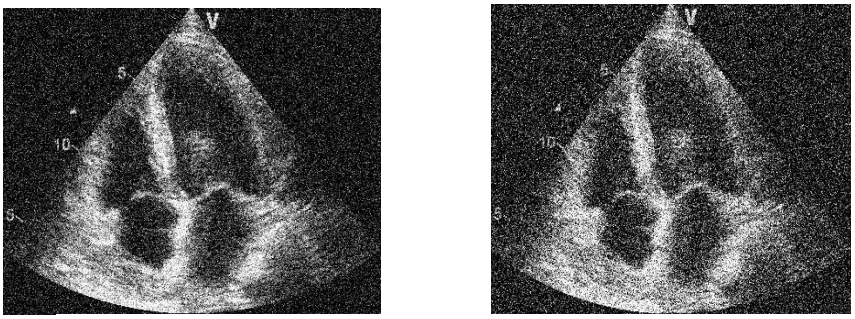
Fig 5. 10 Stability scores for rotation changes

Effect of Noise

Since it is very easy to generate the noises in ultrasound detecting, we compare the stability evolution of different noise levels in all the detection methods. In our test, the Gaussian white noise with different standard deviations is applied to simulate the noises in echocardiogram videos. Figure 5.11 visualizes the effect of different sampling noise levels in our test dataset. With the increasing level of noise, the cardiac structural details disappear gradually.



(a) $\sigma_{noise} = 0.01$



(c) $\sigma_{noise} = 0.06$

Fig 5. 11 The 4A viewpoint image with different levels of noise.

The test result with different standard deviations σ_{noise} (from 0.01 to 0.12) is shown in Figure 5.12. The stability score decreases with the noise increasing for all methods. For dense features, our method demonstrates higher robustness than the dense Gabor method.

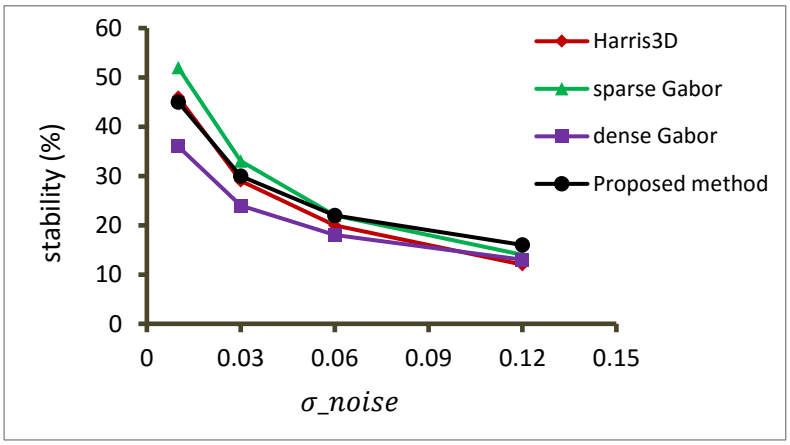
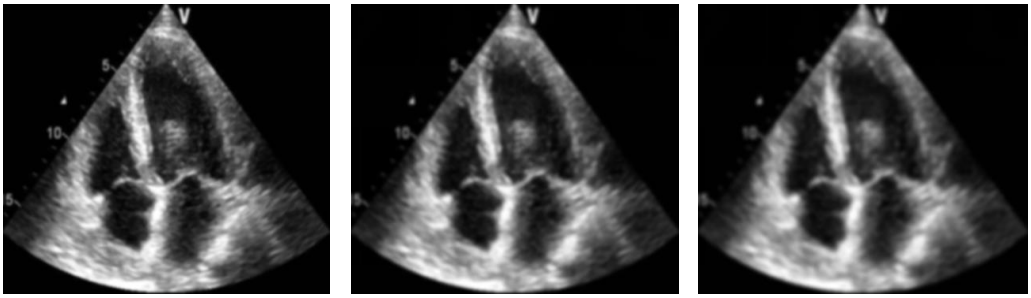


Fig 5. 12 The stability score with increasing the level of noise.

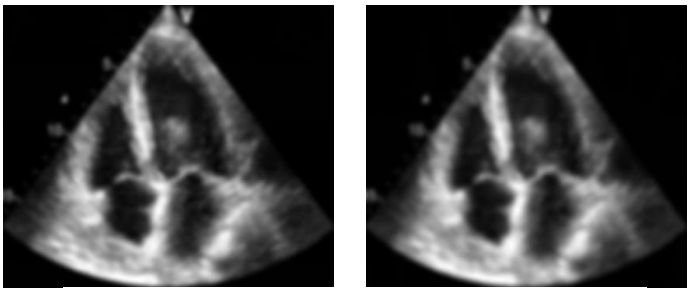
In terms of the experimental results, the anti-noise property of undergoing multi-scale filtering is better than single scale filtering result (dense Gabor feature). But the single scale filtering process can still reduce the noise to a certain degree, which might be the reason that its stability score is close to others.

Blurring

The cardiac structures are constantly in motion when the ultrasound transducer samples the image of the heart during the detecting period. Except the noise, the image blurring derived from heart motion is another factor to impair the detecting result. In our test as illustrated in Figure 5.13, we blur the echocardiogram video sequence by means of Gaussian smooth kernel with different standard deviations σ_{blur} (from 1.5 to 3.5). With the increasing of the standard deviation, the boundary detail in the echocardiogram image is blurred just like a kind of motion stage caught instantaneously by the ultrasound transducer.



(a) $\sigma_{blur} = 1.5$ (b) $\sigma_{blur} = 2$ (c) $\sigma_{blur} = 2.5$



(d) $\sigma_{blur} = 3$ (e) $\sigma_{blur} = 3.5$

Fig 5. 13 The blurring image with different levels of Gaussian smooth.

Figure 5.14 presents the change of stability versus blurring level. The dense Gabor method is clearly more sensitive to this kind of transformation. This can be explained by the lack of relative stable discriminative criteria, but merely depend on the filter respond based on spatial and temporal scales respectively. Comparatively speaking, Hessian saliency and Harris methods outperform Gabor method for the feature detecting in the blurring situation.

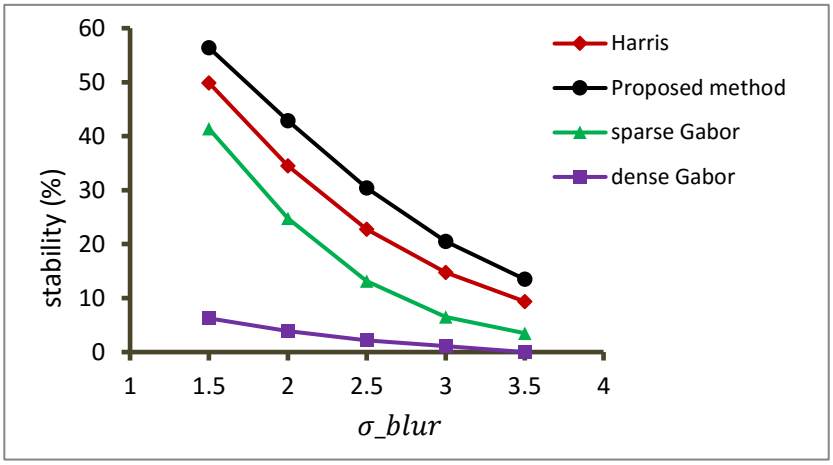


Fig 5. 14 The stability score with the increasing level of Gassian smooth kernel.

Intensity variance

In the process of ultrasound detecting, the intensity of the echocardiogram image has slight changes owing to different sonographers or ultrasound equipment settings. In order to show the stability of the detected features, we decrease the illuminance in different degrees. Figure 5.15 shows the results for the light changes by using all detecting methods. The results are better than the other changes. All methods have nearly horizontal stability curves, showing high performance of invariance to the intensity variance.

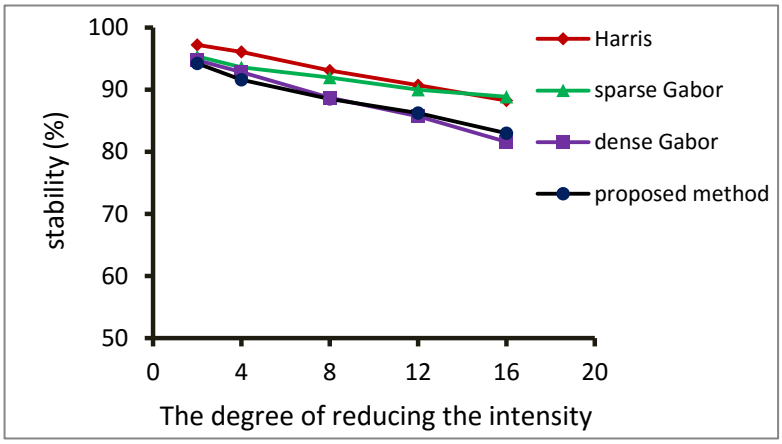


Fig 5. 15 The stability score with deceasing the light of the echocardiogram video sequence.

In general, since our method extracts a dense set of features in the echocardiogram video sequence, the total number of features in each randomly sampled sequence is still higher than what is obtained by the other methods except for dense Gabor method. The dense Gabor feature approach is more sensitive to the noise and blurring variances, with feature points spreading over the whole image, which cannot highlight the salient structural features of the heart. For the echocardiogram video recognition, the dense features can represent more information including spatial distribution, texture character and motion details than sparse features. The Harris3D and sparse Gabor methods show good robustness to the rotation and the intensity variance, but the number of features is too sparse to reflect the majority details of the whole echocardiogram video.

5.5 Summary of the feature point detection

In summary, we develop a 3D KAZE method to detect echocardiogram video features. Based on building a nonlinear scale space, we detect feature points roughly by searching the local maximum determinant of Hessian matrix. And then we use a coarse-to-fine strategy to discard the unreal feature points in each scale spaces. In

addition, we compare our method results with Harris corner and Gabor detectors and evaluate the feasibility and stability of our method in some transforming situations for echocardiogram image detection. Experimental results show that 3D KAZE features mostly accurately locate on the cardiac structures of boundaries and can depict echocardiogram features densely, which shows good performance with varying geometric and photometric transformations.

6. Feature descriptor in acceleration field

Recent studies ([82], [85], [96]) have shown that local features and their descriptors have significant impact on the performance of the video recognition. As mentioned in Section 2.3.2, many descriptors have been developed to classify different datasets from 2D image to 3D spatial and temporal space. Each descriptor shows good performance in the corresponding research. In our research, we need to consider that how to encode important information of feature points to represent the characters of the whole video properly. But what is the important information in echocardiogram videos and how to represent it is the task to be solved in the following research.

The interesting or important information for echocardiogram video sequence refers to distinguishing features that can reflect the cardiac structural information or the cardiac static appearance as well as motion stage. In this chapter, we focus on proposing a new efficient descriptor to represent motion information of local features and compare it with other motion descriptors. The method should not only be efficient to compute (unlike 3DSIFT and HOG3D with high dimensions), but also reflects the periodic relaxation and contraction of the heart.

6.1 Acceleration field descriptor

As we all know, velocity, as a physical vector can be used to reflect motion status of an object, that are the reasons that optical flow is widely used in action recognition and that hence HOF descriptor becomes popular in encoding the motion information. For an echo video, the motion of the heart is derived from periodic systolic and diastolic functions of myocardium and varies instantaneously in different stages. In a cardiac circle, the myocardium has discriminative stress states

in different structures，which consequently presents different motion states in the echo video. In addition, cardiac motion includes some non-functional movement as well (e.g., translation and rotation). Our aim is to mine the deeper motion information from the stress state to get motion details from cardiac structures.

6.1.1 Acceleration field

Acceleration, in physics, is the rate of change of the velocity, which reflects the stress state of an object. In cardiac circle, the structural motion is caused by the variation of stress state. Firstly, we use the classical Horn-Schunk[90] method to calculate the dense optical flow velocity responses (v_x, v_y). Mathematically, acceleration is defined as the derivative of velocity, which can be approximated by using discrete difference between two sequential frames in optical flow, shown as the formula:

$$a(x, y, t) = \frac{dv}{dt} = v(x, y, t + 1) - v(x, y, t) \tag{6.1}$$

Similar as the velocity field, the acceleration field is composed by horizontal and vertical components (a_x, a_y), which is a vector quantity, it has magnitude and direction as formulated by

$$mag(x, y) = \sqrt{a_x^2 + a_y^2}$$

$$\theta(x, y) = tan^{-1}(\frac{a_y}{a_x}) \tag{6.2}$$

We here develop a new descriptor utilising a similar histogram of orientation based method voting with mag as in HOG and HOF descriptor. However, instead of using the gradients (L_x, L_y) and velocity (v_x, v_y), we use acceleration (a_x, a_y), which reflects the myocardium stress state. In order to be different from velocity

field, we label Eq. (6.1) as acceleration field. The descriptor can be viewed as the histogram of acceleration (HoA) accordingly.

Figure 6.1 shows the comparison of magnitude images in optical flow field and acceleration field. We sample a part of continuous slices randomly during a cardiac circle in a PSAP echocardiogram video. By enlarging the same area separately in optical flow and acceleration field as demonstrated with blue bounding boxes in each corresponding figure, the difference between two magnitude images can be observed clearly.

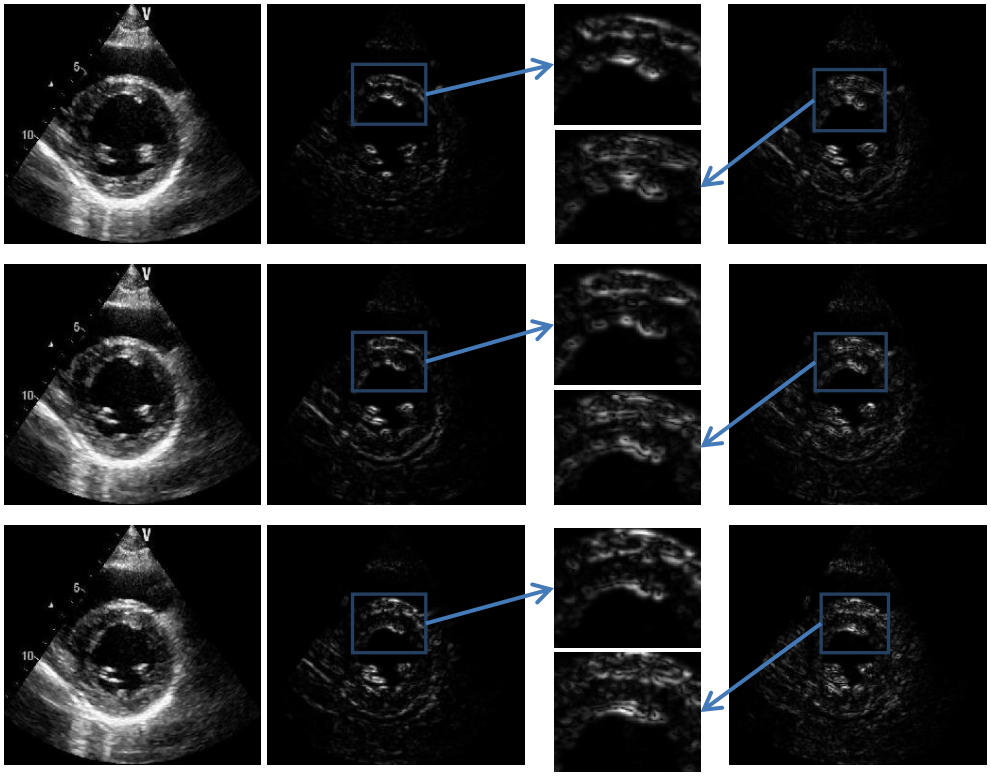


Fig 6. 1 The comparison of magnitude images in optical flow field and acceleration field during the systolic process. The first column shows three consecutive slices. The second and the last columns correspond to the magnitude of velocity and acceleration respectively. The areas inside blue bounding boxes are arranged in the third column after enlarged.

We extract the acceleration field on the position of feature points in each slice and illustrate it using the pink line in Figure 6.2. The arrows point to the acceleration direction of the corresponding feature and the length represents the magnitude of

acceleration. In the systolic period (as shown in the first rows), the myocardium of the LV intensifies the systolic stress state and the LV cavity gets smaller and smaller. Accordingly, in the diastolic process, the acceleration field around the LV shows a centripetal decreasing, as illustrated in the second row of Figure 6.2.

The slices in the bottom show similar changes with the first one. The changing progress of the LV acceleration field shows the consistence with the stress state of the myocardium. In our research, the acceleration field is calculated around feature points detected in Chapter 3 instead of all pixels, which can avoid the influence of noisy points and improve efficiency.

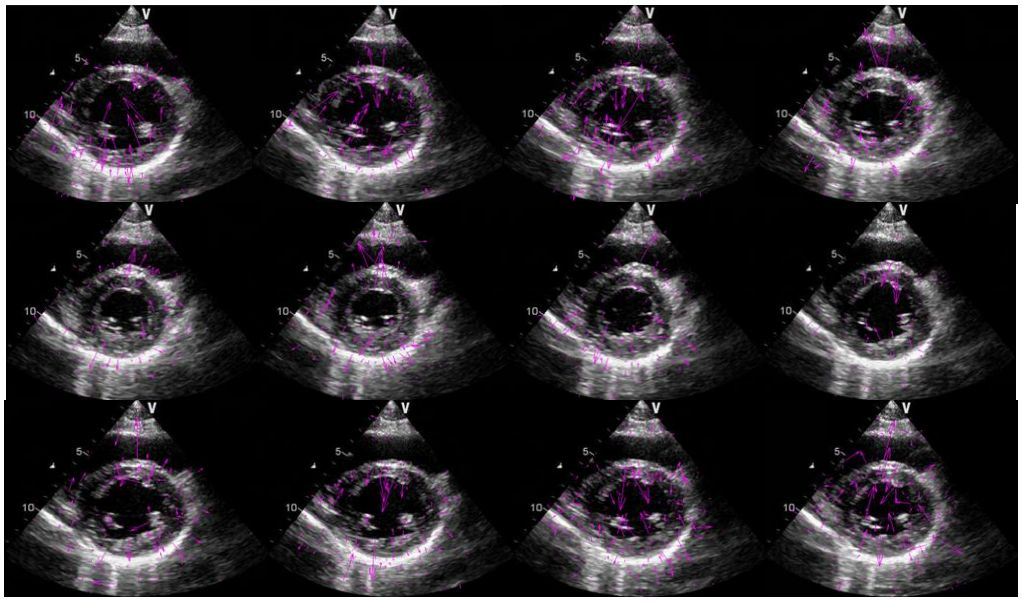


Fig 6. 2 The acceleration field in the cardiac circle. The first row shows the cardiac systolic process. The heart starts to relax from the end-systolic in the second row and then begins to a new round of contraction shown in the last row.

6.1.2 HOA descriptor

The following content explains how to extract HOA (histogram of acceleration) descriptor to encode motion information. Lowe [132] proposes to subdivide local neighbourhoods into parts and compute histograms for each of these parts

separately. As different parts are defined in terms of relative position with respect to the centre of local features, the coarse positional information is preserved in the descriptor. This idea is extended for representing motion events in [84,85,86]. For a spatial-temporal sequence, a set of sub-blocks around the features are divided and represented separately by using the histogram-based representations. The final descriptor is formed by combining all histograms of sub-blocks and shows good performance on the problem of recognizing human actions.

In our research, we extend this idea into the HOA descriptor. Based on the acceleration field, the magnitude and orientation around feature points on each slice are computed by Eq. (6.2). The range of orientation is from 0 to 2π, which is divided into S bins equally (i.e. orientated histograms) in our research. For a sub-volume of echocardiogram video, it can be divided into a set of blocks with size of $M_b \times M_b \times N_b$. For each block, the orientated histograms are voted with weighting based on acceleration magnitudes. The final descriptor is formed by concatenating $m \times m \times n$ adjacent block histograms as illustrated in Figure 6.3. Similar to the descriptor in [86], in order to reflect the position correlation, each block can be recorded many times by the neighbouring descriptor except for the block on the borders of the video volume.

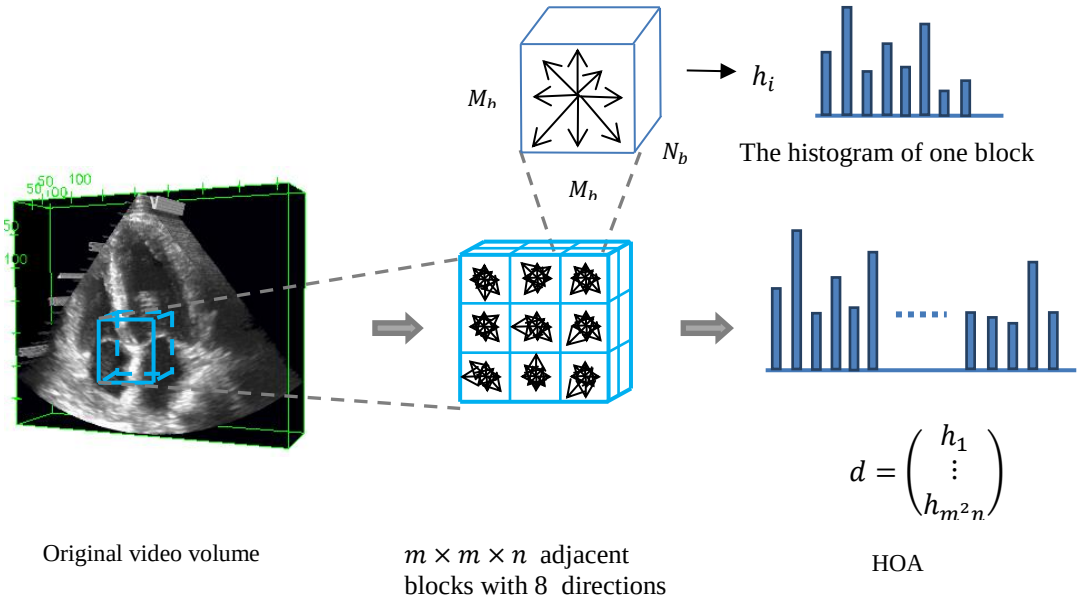


Fig 6. 3 Illustration of the HOA descriptor: the echocardiogram video is divided into a set of blocks; the magnitude of acceleration in each block is calculated, and then a histogram of the magnitude is generated as the descriptor of this block. HOA descriptor corresponding to a certain region is generated by concatenating all orientated histograms of neighbouring blocks.

6.2 HOA descriptor testing and comparing

In this section, since we introduce our HoA descriptor method into application to echocardiogram video classification, we will compare this approach with a number of existing motion information descriptors.

6.2.1 Normalization of acceleration ─ relative acceleration

Before any further processing, we perform normalisation of acceleration first. Normalization is very important to the performance. It can normalize the data into a uniform range for further analyzing and comparing. For HoA descriptor, we normalize all orientated histograms of neighbouring blocks with $L_2\text{-}norm$ when concatenating them into a final descriptor. Let q be the original vector with N dimensions. The normalization vector $'$ is obtained through:

$$q' = \frac{q_i}{\|q\|_2} \qquad (i = 1 \ldots N) \tag{6.3}$$

It is confirmed experimentally that normalization can improve the classification accuracy slightly, which is shown in Table 6.1.

Normalization	A2C	A3C	A4C	A5C	PLA	PSAA	PSAM	PSAP	Mean
No	90.3%	87%	89.7%	55%	97.5%	94.7%	64.6%	81%	85.4%
Yes	90.3%	87%	91.4%	60%	97.5%	96.5%	68.8%	85.7%	86.6%

Table 6. 1 The classification accuracy comparison of before and after normalization

6.2.2 Determination of 3D sub-block information

There are three parameters that can be used to adjust the complexity of our descriptor: the number of orientations S in the histograms, and the number of blocks ($m \times m \times n$ i.e. m by m blocks in the spatial domain and n blocks in the temporal domain) for one descriptor and the size of each block $(M_b \times M_b \times N_b)$. Hence the dimension of the resulting *descriptor v*ector is $S \times m \times m \times n$.

In our case, we use S bins ($S = \{4,8,16,24,32\}$) to represent the oriented histogram respectively and $3 \times 3 \times 2$, $2 \times 2 \times 2$, $4 \times 4 \times 2$ blocks in the spatial and temporal domains for the descriptor. Figure 6.4 shows the efficiency of different parameters in generating the descriptor. Because the dimension increases with the increase of the number of oriented bins, the efficiency of generating the descriptor decreases correspondingly. Figure 6.5 illustrates the accuracy comparison of three block settings with eight oriented histograms in echocardiogram viewpoints classification. We evaluate the accuracy and efficiency trade-off for different parameter pairs and use 8 oriented histograms and $4 \times 4 \times 2$

blocks to form a 256-dimensions (8 $orientations \times 32\ blocks$) descriptor in our following implementations.

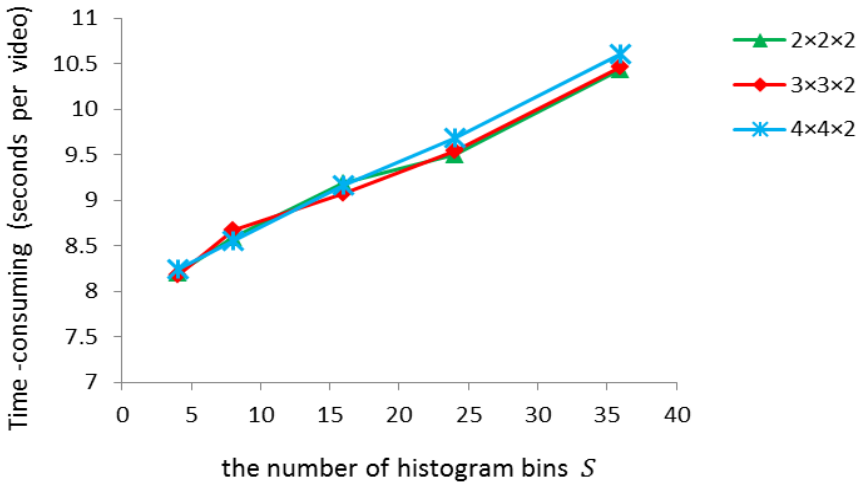


Fig 6. 4 The efficiency comparison of the number of oriented histogram and the number of spatial-temporal blocks in the process of generating HOA descriptor.

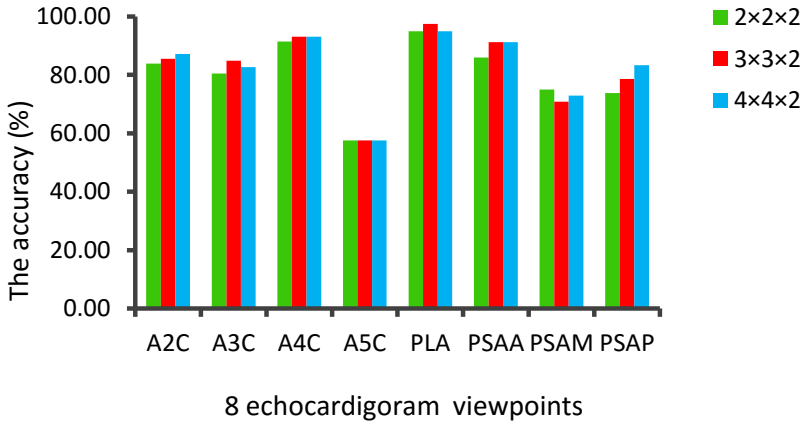


Fig 6. 5 The accuracy comparison of different block settings ($m \times m \times n$) in echocardiogram viewpoints classification.

Based on the settings as mentioned above, each echocardiogram video can be represented by a set of descriptors. For a video with the size of $341pixels \times 415piels \times 26frames$ in our dataset, the number of descriptors varies with the size of block ($M_b \times M_b \times N_b$) . The smaller the size is, the larger the number of

descriptors is. Table 6.2 shows the effect of the block size on the mean average precision (mAP) of the echocardiogram viewpoints classification. For each viewpoint in our dataset, the classification accuracy is different by changing the block size, which is illustrated in Figure 6.6. Considering the accuracy of all viewpoints recognition, we use the size of $12 \times 12 \times 6$ as the block size in our experiment. So each echocardiogram video can be represented into a final vector with the size of 2325 descriptor×256 dimension. The different block sizes and corresponding characteristics are illustrated in Table 6.2.

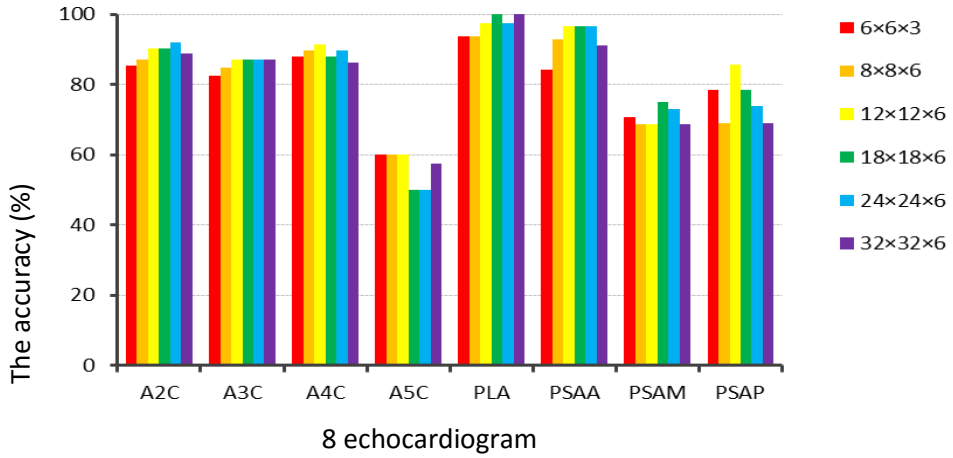


Fig 6. 6 The accuracy comparison of 8 echocardiogram viewpoints based on different block size settings.

histogram	8					
block count per descriptor	4×4×2					
descriptor dimensionality	256					
block size	6×6×3	8×8×6	12×12×6	18×18×6	24×24×6	32×32×6
descriptor count per video	24486	5616	2325	900	462	189
mean AP	82.2	82.9	86.6	85.7	85	83.6

Table 6. 2 The comparison of different sizes of block. The last row shows the mean average precision in echocardiogram video classification corresponding to different block sizes in the fourth row.

6.2.3 The comparison with other descriptors

Table 6.3 shows the comparison of HOA descriptor with other descriptors. In our implementation, the spatial gradient descriptor (HOG and 3DSIFT) and the motion information descriptor (HOF and MBH) are applied to encode the echocardiogram video as well. We give results obtained from each descriptor separately, but also for possible combination between HOG and HOF descriptors.

	spatial gradient descriptors		motional information descriptors			combination
descriptor	HOG	3DSIFT	MBH	HOF	HOA	HOG/HOF
mAP (%)	85.65	84.72	83.56	85.65	86.57	86.34

Table 6. 3 The comparison with other descriptors in echocardiogram classification.

For HOG, HOF and MBH descriptors, we use the method introduced in [86] to implement. But the difference is about the descriptor extraction. Uijlings et al formulate the gradient and dense optical flow at a sampling rate of the entire image region. In our implementation, we extract the gradient and motion information in the neighborhood around the local features based on our feature detector. Since all the extracted descriptors correspond to a set of specific characters or structures instead of distributing in the whole image domain uniformly. It can not only reduce the influence of the noise but also highlight the feature characteristics. The parameters about the block and histogram are the same as HOA descriptor settings. Accordingly, the video is represented into a final vector with the size of 2325×256. In addition, for HOG/HOF descriptor, 4-bin oriented gradient histograms and 4-bin oriented optical flow histograms are computed and concatenated to form the final vector.

We can observe from Table 6.3 that our proposed descriptor, i.e. using the acceleration field around the feature points to describe the motion information, outperforms the histogram of optical flow by 1% and motion boundary histograms by 3% for the recognition of echocardiogram videos. Compared with spatial gradient descriptors, it shows better performance as well. The detailed results for eight viewpoints classification are illustrated in Figure 6.7.

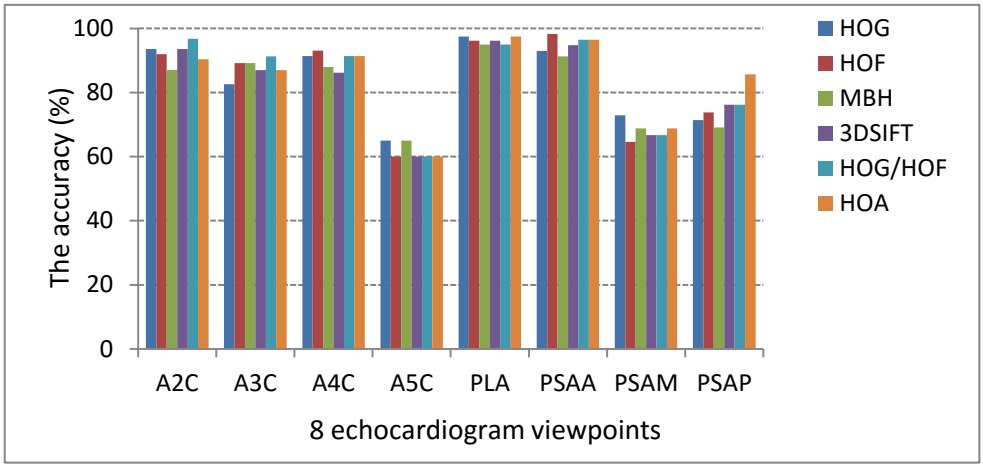


Fig 6. 7 The accuracy comparison of different descriptors on the 8 viewpoints classification of echocardiogram video

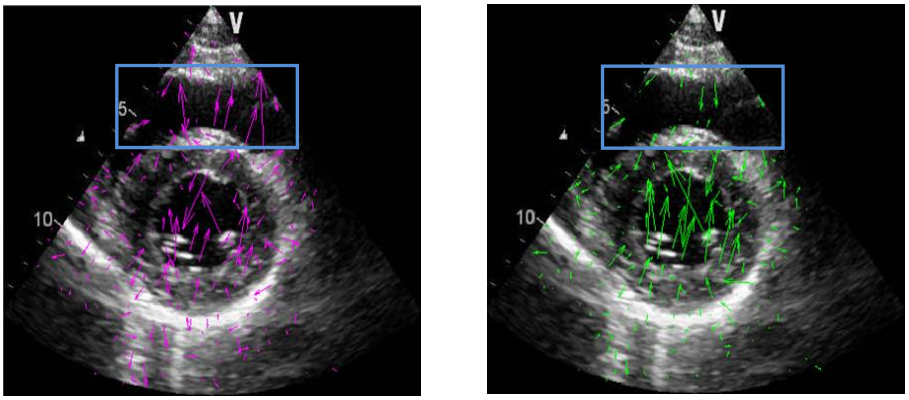
6.3 Discussion of descriptors

For spatial gradient descriptor, HOG shows better performance than 3DSIFT as a whole. This can be explained by the fact that the resolution between spatial and temporal domain is different, which results in a certain degree of errors when calculating the magnitude and orientation of 3DSIFT descriptor.

For motional information descriptors, all of them are based on the optical flow (v_x, v_y). The motion boundary histogram (MBH) description separates the optical flow field into x and y components and computes HOG descriptor separately. Since it represents the gradient of the optical flow, constant motion information can be suppressed. Actually, for the echocardiogram video, all structures are in motion

including functional (systolic and diastolic motion) and non-functional movements (caused by structure traction and breathing). It changes constantly in the whole cardiac circle instead of constant motion like camera movement in human action scene in [82]. So MBH cannot play a significant role in the echocardiogram video recognition. HOF descriptor shows better performance than MBH in the echocardiogram video classification. It reflects instantaneous motion state by calculating the velocity of the cardiac structures.

Unlike HOF descriptor, HOA descriptor reflects the motion state by recording the changing of velocity. Furthermore, according to Newton's Second Law, acceleration field can embody the stress state of the corresponding cardiac structures. Figure 6.8 shows the acceleration and velocity fields on one of the slides in the cardiac systolic process. The region inside the blue bounding boxes corresponds to the cardiac RV (right ventricle). It contracts synchronously with the LV. The myocardium systolic trend of RV in acceleration field is more prominent than that in the velocity field.



acceleration velocity

Fig 6. 8 The illustration of acceleration and velocity field in a slide during the cardiac contraction period.

Among eight cardiac viewpoints, the classification accuracy of A5C viewpoint is lower than others (as shown in Figure 6.7). It can be explained that the silhouette of aortic root (AO) surrounded by four chambers is not obvious. The motion state of aortic valve (AV) cannot be captured easily and clearly because of its dim moving in comparison with other parts of structural motions, as illustrated in Figure 6.9. In addition, its shape has great similarity with A4C, which can be shown in Section 3.2.1 at the introduction of A4C and A5C.

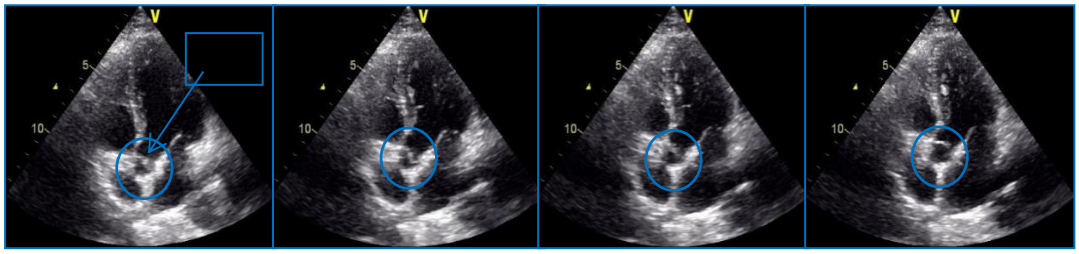


Fig 6. 9 The motion state of AV in four consecutive slides of an A5C video. The blue circle area corresponds to cardiac aortic root and aortic valve.

6.4 Summary of HOA descriptor

This chapter proposes a novel descriptor based on the acceleration field and evaluates its performance based on viewpoint recognition in echocardiogram videos. We analyze the feasibility of HOA based on the accuracy of classification of different viewpoints from two different perspectives (i.e. the variation of velocity and cardiac stress state) and analyze the cause of low accuracy of some viewpoint recognition. In addition, we compare HOA descriptor with other popular descriptors including spatial descriptors, motional information descriptors and the combination of both in echocardiogram classification. Finally, the experiment shows that the proposed descriptor has better performance for the echocardiogram video classification.

7. Echo Classification based on 2D and 3D KAZE feature points

The classification procedure in this study employs a) feature detection using KAZA (section 4); b) feature description based on acceleration field (section 6); c) encoding features using Fisher Vector approach, and d) classification using SVM. In this chapter, we evaluate the performance of different feature representations (including Fisher vector and Bag-of-Word) and classification strategies when using SVM to recognize echocardiogram videos.

7.1 Video representation

A video can be represented by a set of feature descriptors after features are detected and described. This kind of feature descriptor derived from original video sequences can be viewed as low-level features. It is likely to contain a large number of redundant information and causes time-consuming in the subsequent training process. In order to improve the robustness of feature expression and be suitable for classification, feature descriptors should be encoded into higher level of discriminative presentation. Therefore, encoding feature is necessary to convert descriptor into another kind of vector when classifying.

As mentioned in Section 2.3.3, Bag-of-Words is the most popular method for encoding features. Fisher vector (FV), as an extension of BOW, is confirmed recently to bring large improvements in terms of accuracy for image classification ([133], [93], [134], [135]). In our implementation, we encode descriptors using FV and BoW respectively.

7.1.1 Fisher vector

FV encoding assumes that descriptors are generated by a GMM (Gaussian Mixture Model) model with diagonal covariance matrices. The GMM model of K Gaussians,

which can be viewed as 'codebook', is first obtained from a training set. Once the Gaussian model $(\omega_i, \mu_i, \sigma_i)$ is learned, the FV representation $(U_{\mu,i}^X, V_{\mu,i}^X)$ can be obtained. It aggregates much raw information (described in each descriptor x_t) into a high dimensional vector and encodes some additional distribution information of the Gaussian distribution (e.g. mean μ_i and standard deviation σ_i). So it has been reported to consistently improve the performance in the image classification.

In the training stage, we sample N HOA feature descriptors in all the datasets to generate the GMM model $(\omega_i, \mu_i, \sigma_i)$. In our experiments, we use VLFeat library [136] to implement Fisher vector encoding.

In order to evaluate the parameter settings, i.e. the number of sampled descriptors N and the size of the codebook K, we set K from 16 to 512 and N from $40k$ to $160k$. The results are illustrated in Figure 7.1.

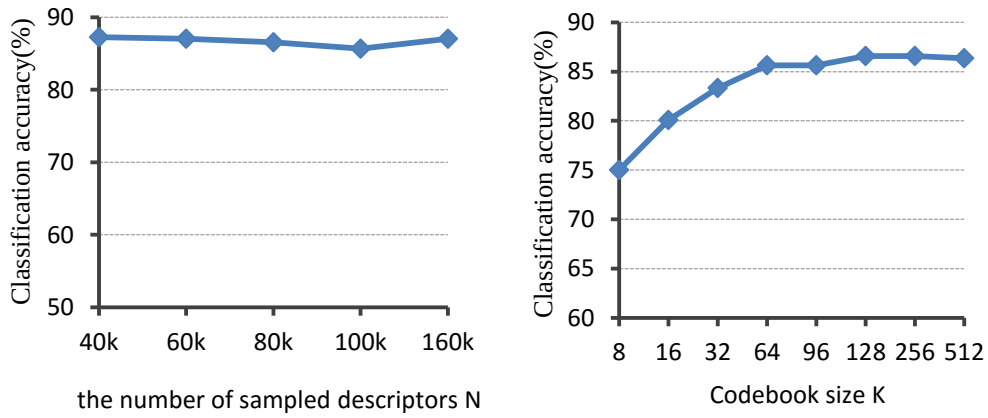


Fig 7. 1 Performance of classification with the variance of the number of sampled descriptors N and the code words K.

It is clearly showing that the number of sampled descriptors N in FV representation has little effect on the classification performance. While the variance of the size of codebook shows obvious influence. In our experiment, we consider

the accuracy and efficiency trade-off and fix the number of sampled descriptors to $N = 80k$ and the codebook size to $K = 128$.

7.1.2 Bag-of-features approach

In our research, a standard BoW baseline, which can be viewed as hard assignment, is applied to encode features. It is the classical way of transforming a set of local visual descriptors into a single fixed-length vector by using a k-means visual vocabulary (codebook) and assigning local descriptors to the nearest cluster, which can be illustrated in Figure 7.2.

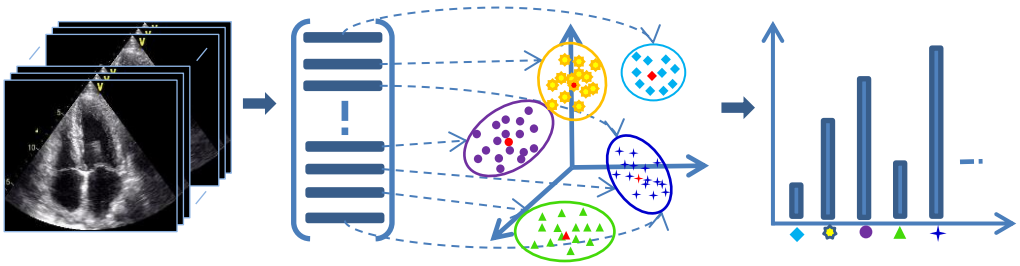


Fig 7. 2 The illustration of k-means clustering for echocardiogram video classification. Original video is detected and described into a set of D-dimensional feature descriptors, which is the raw feature representation. And then by using k-means clustering, feature descriptors are transformed into a k-dimensional histogram vector. Finally, all of the echocardiogram videos are represented into a histogram matrix. Rows of the matrix correspond to k clusterings and columns correspond to the number of videos.

In order to add some spatial information in the BoW encoding method, we use a standard BoW approach with the following modifications. First, instead of hard assigning each local feature to its nearest clustering centre, soft assignment to the nearest n centres used by Chen Sun et al. [137] is applied to increase clustering information in the K–bin histogram. Next, we use spatial-temporal pyramid to encode spatial and temporal structures. As illustrated in Figure 7.3, we apply the spatial Pyramid Matching (SPM) as described in [78] to divide each

echocardiogram video into a set of sub-volumes. The K–bin histograms for all sub-volumes are computed and then concatenated into a final spatial-temporal vector.

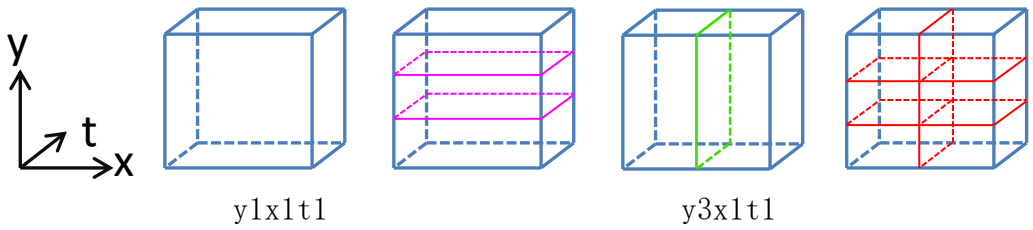


Fig 7. 3 The illustrations of a set of sub-volumes for each video in BoW.

Additionally, another method for constructing visual dictionary is Sparse coding (SC) [98], which models feature vector as a sparse linear combination of a set of basic elements (also called dictionary). The number of basic elements can be defined artificially. As similar to k-means, we use spatial pyramid to divide each video. For each feature descriptor in corresponding sub-volume, a set of sparse codes are generated followed by the max pooling and then concatenated into the final vector. The coding process is depicted in Figure 7.4.

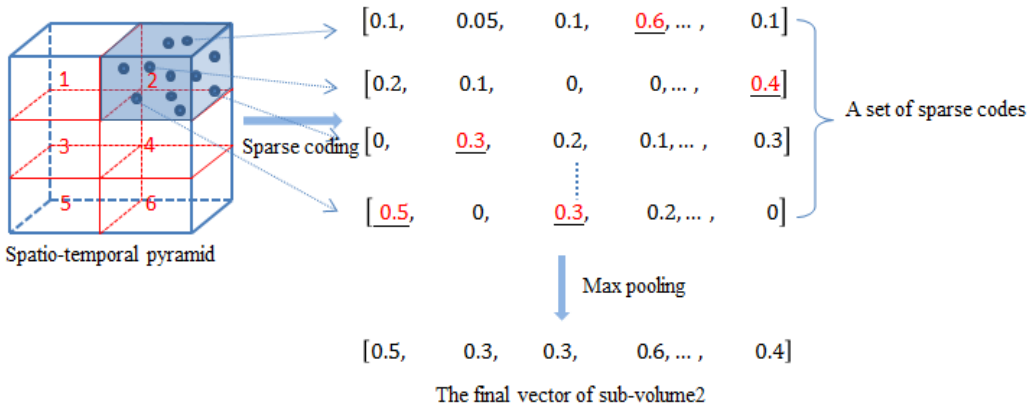


Fig 7. 4 The pipeline of sparse coding. We divided echocardiogram video into a set of sub-volumes. For each feature descriptor in corresponding sub-volume, a set of sparse codes are generated followed by the max pooling. The representations for each sub-volume are concatenated into the final vector of the video.

7.1.3 The comparison of FV and BoW

The basic idea of both FV and BoW is to transfer feature set into a fixed dimension vector, from which the distribution in the original feature space can be reconstructed approximately. In this section, we compare the FV to the BoW (using k-means and sparse coding methods to generate codebook respectively). In training stage, $80k$ feature descriptors are randomly sampled from all features in our video datasets.

Based on experimental results, we set codebook size as $K = 256$ for k-means and 1024 for sparse coding. In k-means, the number of nearest centers is set with $n = 10$. A video is divided into 12 sub-volumes as illustrated in Figure 7.3 (with 3 sub-volumes in y direction, 2 along x direction and 6 sub-volumes from both x and y directions, as well as one original volume). The final representation has 3072 (256×12) dimensions and is l_2-normalized. For sparse coding, the codebook with the size of 1024 shows better performance in our experiments. In coding stage, the final vector is of 12288 (1024$\times$ 12) dimensions.

Figure 7.5 presents the performance of comparison of feature encoding approaches between FV and BoW. On the base of the same size of the codebook, the Fisher vector shows better performance than BoW approach, e.g. for $K = 128$ (the default setting of FV encoding in our experiments), FV improves around 4% performance than BoW with sparse coding, above 20% than standard BoW and 2.3% than modified BoW. This result is consistent with the report in [93] on image classification. In addition, we note that our modified BoW improves the performance significantly comparing with standard BoW method.

Additionally, we consider combining FV with the Spatial Pyramid to highlight the spatial distribution information of features, as reported in [93,104]. Being similar to the modified BoW, we use a very coarse spatial pyramid to divide each

echocardiogram video into 6 sub-volumes: one FV for the whole image, three FV in y direction corresponding to the top, middle and bottom regions of the image, and two FV in x direction corresponding to the left and right regions respectively, which is illustrated in Figure 7.3. The final vector is obtained by concatenating all representations of sub-volumes. The main challenge is that the high dimensionality is expensive in classification process even by using linear SVM. Considering the accuracy and efficiency of trade-off, we decrease the number of Gaussians K from 128 to 32. Accordingly, the final FV representation becomes $12DK$ ($6 \times 2 \times 256 \times 32 = 98304$) dimensional. The final classification results are illustrated in Table 7.4. Fisher Vector is the default representation method in our experiments unless noted otherwise.

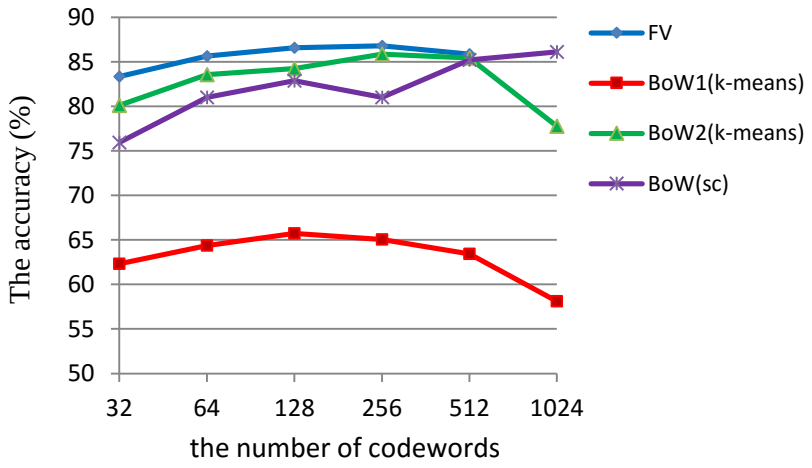


Fig 7. 5 Comparison of performance by using different encoding methods in echocardiogram video classification. For BoW, we use k-means and sparse coding (sc) methods respectively to generate codebook. When using k-means, BoW1 corresponds to the standard baseline, while BoW2 corresponds to the modified method.

7.2 Multi-class SVM

Support Vector Machines (SVM) with different kernel functions is originally designed for binary classification. In our research, all classification results are

derived from SVM. It includes two processes: training and testing. The classifier generated from training data is applied to classify testing data into corresponding classes.

7.2.1 one-versus-one vs. one-versus-all

SVM can be applied to multi-class classification as well by decomposing the multi-class problem into a set of two-class problems. There are two popular decomposition methods: one-versus-one (OVO) and one-versus-all (OVA). For n classes, in one-versus-one method, $n(n-1)/2$ classifiers are trained with data from any two of all classes. The prediction for each tested video is implemented according to the maximum voting, where each SVM votes for one class. While n SVM classifiers are constructed in one-versus-all method. In the training process, the ith class is set with positive label, while others with negative label. Each SVM is trained to distinguish the data of one class from all the remaining classes. And then, the tested video belongs to the class with the greatest probability.

Which one has better performance in classification has no unified conclusion. The one-versus-one strategy is substantially faster than one-versus-all method[138], which is confirmed from our experiment results (as shown in Table 7.1). OVO shows better recognition performance in[139,140], while [141,142] report that OVA performs better in action recognition, even on large scale datasets as well. Which is more suitable for echocardiogram video classification, there is no report for this question so far yet.

In our experiment, we use multiclass SVM with OVO and OVA strategies separately to classify echocardiogram videos based on the same parameters and compare the performance from the accuracy of classification. The detailed comparison between two strategies is shown in Table 7.1. Our experiments show slightly better recognition accuracies when using OVA on our histogram of acceleration descriptors in echocardiogram video classification.

method		Mean accuracy (%)	Seconds per video (s)
Linear kernel	OVO	86.1	138
	OVA	88	436
RBF kernel	OVO	85.2	259
	OVA	86.1	928

Table 7. 1 The comparison of performance using different SVM strategies to classify echocardiogram videos.

7.2.2 Different kernels in SVM

Because different kernel functions have different effects for classification results, we study the influence on echocardiogram video classification from two kernels: linear kernel and Gaussian kernel (RBF). For its implementation, we adopt LIBSVM [140] package with the default parameter settings. According to Table 7.1, both kernels provide good accuracy in classifying. But for efficiency, RBF kernel is more time-consuming than linear kernel. Considering the accuracy and efficiency, we use linear multiclass SVM with one-versus-all strategy in our research.

7.3. Classification of echocardiogram videos

In this Section, we are focus on experiments of echocardiogram video classification. We divide our experiments into two groups: 2D space domain and 3D spatial-temporal domain. For 2D space domain, we extract features (including 2D KAZE and 2D sift features) frame by frame and encode them into Fisher vector before using linear multiclass SVM to classify. About spatial-temporal domain, the experiments are implemented according to several cases shown in Section 7.3.2.

In our experiments, the dataset consists of eight classes including 432 videos altogether. The detailed information about the dataset is illustrated in Chapter 3. The framework of echocardiogram video classification is exhibited in Figure 7.6.

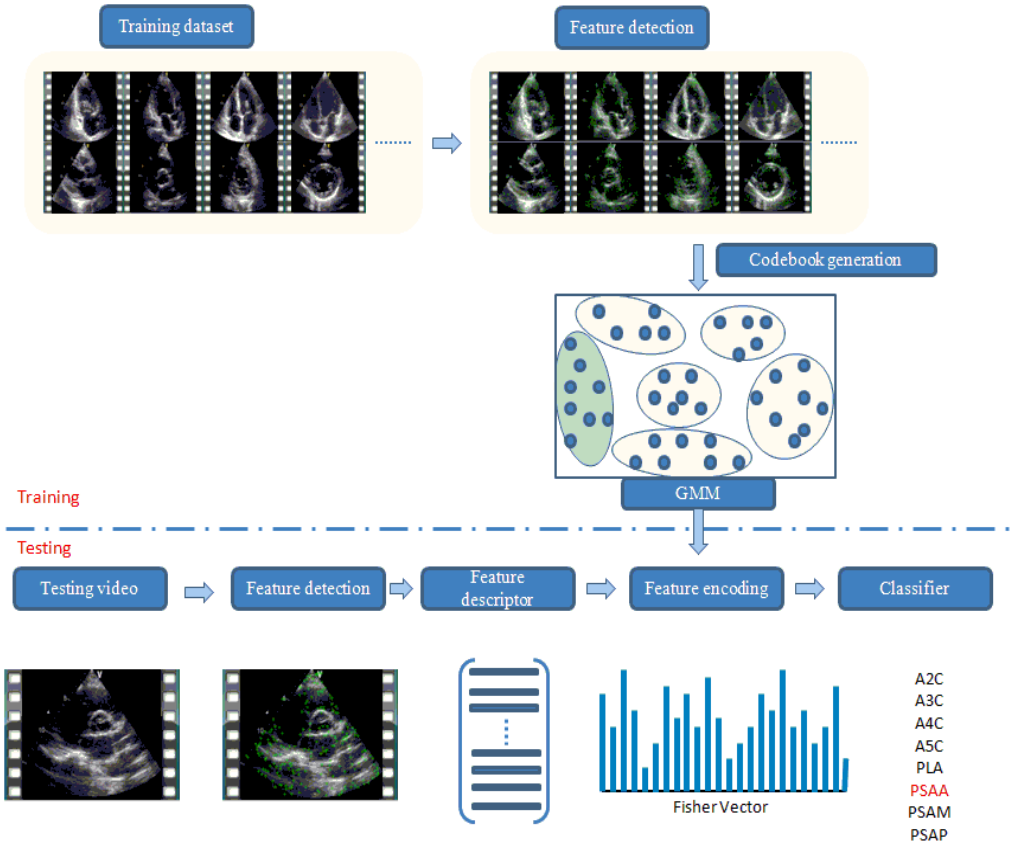


Fig 7. 6 The pipeline of echocardiogram video classification. It mainly includes two stages: training and testing. In training process, codebook can be generated using random training samples. For testing stage, it mainly includes five steps.

In the implementation, unless stated otherwise, all the results, as well as those mentioned above, are reported with FV encoding as well as linear multi-class SVM with one-versus-all strategy. Additionally, we compare our methods with other state-of-the-art in echocardiogram classification.

7.3.1 Echocardiogram image classification in 2D space domain

In our experiment, we introduce 2D SIFT and 2DKAZE technologies separately into echocardiogram classification. A set of SIFT and KAZE features are detected frame by frame as shown in Figure 2.7. For feature representation, Fisher vector is applied to encode all of features. We use VLFeat library [136] and KAZE code package [143] to extract SIFT features and KAZE features respectively with the corresponding default parameter settings. The confusion matrix for our dataset can be illustrated in Figures 7.7 and 7.8.

Fig 7. 7 The confusion matrix of 8 echocardiogram viewpoints classification using 2DSIFT detecting method. The average accuracy is about 83.8%.

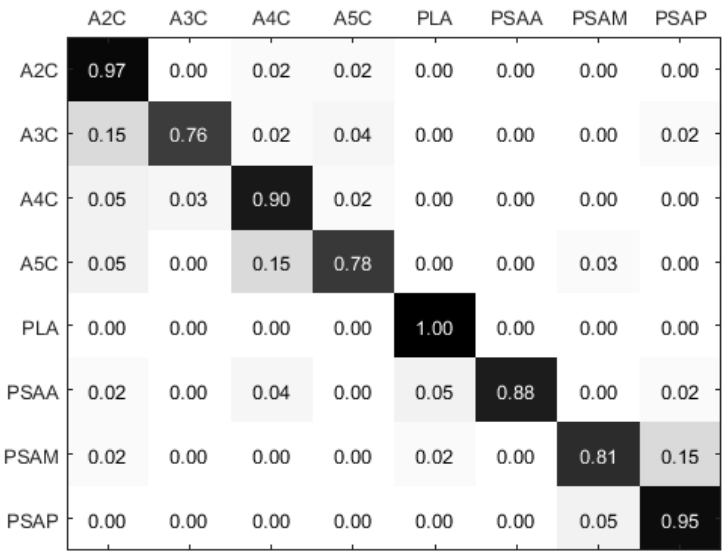


Fig 7. 8 The confusion matrix of 8 echocardiogram viewpoints classification using 2DKAZE detecting method. The average accuracy is about 89.4%.

7.3.2 Echocardiogram video classification in spatial-temporal space

Based on spatial-temporal space classification, we divide experiments into the following situations: 2D KAZE detector combining with the histogram of optical flow, dense optical flow detecting, Cubiod detector with 3D SIFT [78] and 3D KAZE detector with HOA. For encoding features, they are all represented using Fisher vector with $K = 128$.

Since optical flow is the most popular method for describing temporal information in action recognition. In our experiment, firstly we consider to combine optical flow with 2D KAZE feature detecting to depict echocardiogram features. We then calculate the optical flow around the detected feature points and describe it using histogram method as discussed in Section 6.2.3. In addition, the dense optical flow is applied directly to reflect cardiac motion information without

considering KAZE features. The confusion matrices using these two methods to classify echocardiogram videos are illustrated in Figure 7.9 and 7.10 respectively.

Fig 7. 9 The confusion matrix of 8 echocardiogram viewpoints classification using 2DKAZE feature detection combining with optical flow method. The average accuracy is about 84.3%.

Fig 7. 10 The confusion matrix of 8 echocardiogram viewpoints classification using dense optical flow to describe cardiac motion information. The average accuracy is about 79.4%.

In addition, we use the method as mentioned in [78] to implement classification on our dataset, i.e. Cuboid detector and 3DSIFT descriptor are applied to represent original features of echocardiogram videos. BoW with default parameters is adopted to transform the features into final vector. Training and testing processes are the same with our method. In terms of our method, based on 3D KAZE feature detector and histogram of acceleration descriptor, we encode features using Fisher vector as depicted in Section 7.1. The confusion matrixes for echocardiogram videos classification are obtained and stated separately in Figure 7.11 to 7.12.

Fig 7. 11 Confusion matrix using the method in [78]. The average accuracy is about 73.8%.

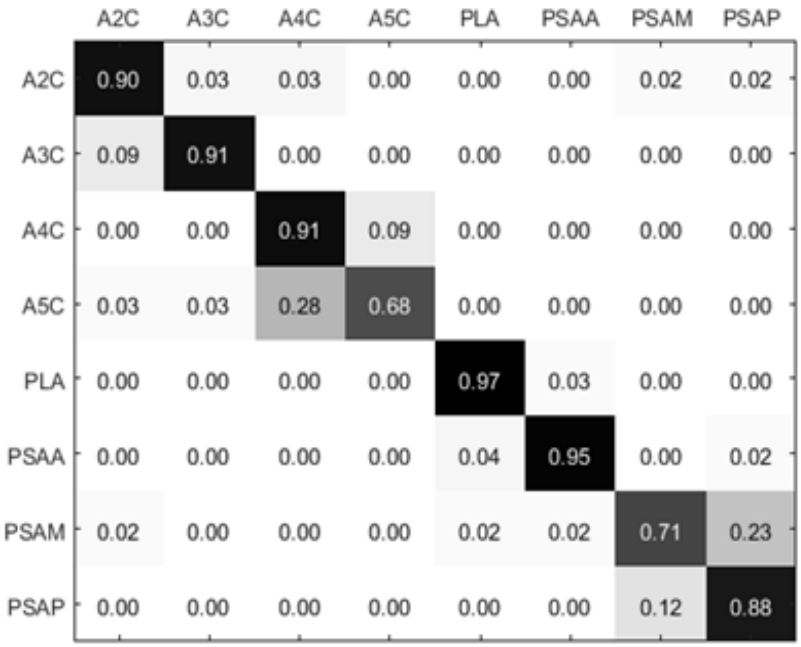


Fig 7. 12 Confusion matrix of our method using 3D KAZE feature detecting and Fisher vector encoding. The average accuracy is about 87.9%.

Table 7.2 shows the results of all classification strategies in our experiments. All the results are derived from the same echocardiogram videos dataset.

Methods		Average accuracy
2D space domain	2D KAZE	89.4%
	2D SIFT	83.8%
Spatial-temporal domain	2D KAZE+Optical flow	84.3%
	Dense Optical flow	79.4%
	Cubiod detector+3D SIFT	73.8%
	3D KAZE detector + HOA	87.9%

Table 7. 2 The illustration of classification results using different methods to implement. All the experiments are based on the same dataset.

7.3.3 Comparison with the state-of-the-art

For echocardiogram video classification, there are many state-of-the-art methods reporting good results. In our method, we extract features by using 3D KAZE

detector and describe these features by applying histogram of acceleration descriptors (HOA). Then, all of the features are encoded into a fixed-length vector to represent corresponding echocardiogram video. Finally, on the base of multi-class SVM classifier, all of the tested videos are classified into corresponding classes. In our experiments, the feature descriptors are fixed with 256-dimensional. The GMMs with $K = 128$ are trained by using $80k$ sampled feature descriptors. Additionally, we use k-means and sparse coding methods respectively to generate codebook as well and combine with spatial pyramid (SP) to add geometric information in the final representations. From the experiment, we apply 256 codewords for k-means and 1024 for sparse coding when generating codebook, and use the same sampled features as done in FV for training. Due to the small dataset in our dataset, we learn linear multi-class SVM by employing the leave-one-out cross validation methodology, i.e. when testing a video clip, the entire dataset exclude test video is used for SVM training.

In addition, our method is also tested on four primary viewpoints including A2C, A4C, PLA and PSA (parasternal short axis including PSAA, PSAM and PSAP). All of these viewpoints are frequently classified in the literature [112, 34, 36, 108, 105, 35]. Table 7.3 shows the results based on Fisher vector representing. A total of these four classes are 346 videos (as stated in Section 2.4). All the classified processes and parameter settings are the same with that in eight viewpoints classification.

Ground Truth(346)	A2C	A4C	PLA	PSA	Accuracy rate(%)
A2C	60	2	0	0	96.8
A4C	1	55	0	2	94.8
PLA	0	0	76	3	96.2
PSA	1	0	1	145	98.6
Average accuracy					97.1

Table 7. 3 Confusion matrix for 4 primary viewpoints of echocardiogram.

Table 7.4 shows the comparison results with other state-of-the-art methods in echocardiogram image or video classification. Note that the mean accuracy of each method is obtained from the corresponding literature.

method		The number of classes	Mean accuracy(%)
Balaji et al. [2014]		3	94.56
Ebadollahi et al. [2004]		10	67.8
Otey et al. [2006]		4	92.7
Aschkenasy et al. [2006]		4	90
Hui W. et al. [2013]		8	98.51
Park et al. [2007]		4	96.3
Zhou et al. [2006]		2	90.2
Roy et al. [2008]		4	97.19
Agarwal D. et al. [2013]		2	98
Balaji et al. [2015]		4	90.7
Kumar, R. et al. [2009]		4\8	98.4\81
Beymer et al. [2008]		4	87.9
Y. Qian et al. [2013]		3\8	90\72
Our method	$BoW^*_{k\text{-}means}$	4\8	94.5 ±0.3\85.6±0.5
	BoW^*_{sparse}	4\8	96±0.3\85.8±0.6
	FV	4\ 8	97.1 ±0.4\86.8±1.1
	FV^*	4\8	96.5±0.5\86.5±0.5

Table 7. 4 Comparison of average accuracy with state-of-the-art methods in the literature. Those marked with * in our method are modified methods by combining with spatial pyramid technology during encoding. We use k-means and sparse coding methods respectively to generate the codebook in BoW, which is illustrated in corresponding subscripts.

7.4 Discussion and Future work

For the feature representation, there are some differences between BoW and FV. The standard BoW uses a hard quantization of feature space by k-means, where all clusters have the same importance and are described by its centroid only. Each

feature corresponds to one closest clustering centre, which leads to inadequate information represented in histograms. It can be confirmed clearly from the low performance in classification. Modified BoW can help to capture more information, which improves the classification performance significantly. FV shows a little better performance than the modification BoW. Since GMM is the generative model for local features in FV, it incorporates more feature information, e.g. mean and standard deviation, in encoding local descriptors (as shown in Appendix A.1). In addition, Fisher vector combining with spatial pyramid technology can obtain good result in our experiment but non-distinctive and expensive in the following classification when dimensions are increasing. This could be caused by the fact that we decrease the size of Gaussians from 128 to 32 for the efficiency. For the efficiency and accuracy of trade-off, we will not consider other models of spatial pyramid in Fisher vector representation. However, this assumption will constitute one component of our future work.

It is very difficult to conclude which one is better between one-versus-one and one-versus-all strategies when using SVM. In our experiment, the latter shows slightly better performance in echocardiogram video classification, however it is more time-consuming than the former. According to the analysis by [138], each approach is better suited for each specific group of tasks and types of images. For problems with few classes, like our eight cardiac viewpoints, the one-versus-all appears to be more accurate. While one-versus-one strategy is more suitable for a very large number of classes, the large number of the size of training dataset will lead large number of trainings and longer processing time. In our research, because the dataset is in a small number (432), one-versus-all is viewed as the preferred method.

For the choice of kernel function, our experiment result confirms that linear kernel can be applied to classify echocardiogram videos. It shows better performance on the accuracy and efficiency than RBF. This could be due to the fact that we implement SVM classification with RBF kernel based on the default parameter settings. It is possible that RBF kernel can provide better accuracy in classification through exploring numerous trials for parameter settings, which will be part of future work.

In our experiments, we use one-versus-all strategy to implement the classification. In practical terms, one echo video is the testing data, all the remaining echo videos are the training data. By using the classifiers trained using training data, the tested video is labeled to the corresponding class with the greatest probability. After all the echo video are classified by repeating this process, we start to a new round of training and testing until the end of the loop. In our experiment, the default value of loop is ten, which can be changed manually. The classification result is derived from ten times averaging with less than two percent difference of precision between two loops. All the classification results (as show in Table 7.2) are obtained separately with the same process as mentioned above. We will calculate the standard deviation of the classification result so as to further compare and evaluate.

7.5 Summary of proposed classification method

In this chapter, we focus on the implementation of classifications for echocardiogram videos. Based on the same dataset collection, we adopt different methods to recognize echocardiogram videos. Firstly, it is experimentally confirmed that KAZE method shows better performance in echocardiogram recognition than SIFT method in 2D spatial domain. Additionally, with regard to

spatial-temporal space methods, the combination of 3D KAZE detector with histogram of acceleration provides higher average accuracy than the others.

However, the 3D form of either KAZE or SIFT does not provide better classification accuracy rates as expected. For example, 2D KAZE method achieves 89.4% mean accuracy, which is 1.5% better than 3D KAZE method. It is highly possible that we didn't align our datasets, i.e. to align them to start at the same phase of a cardiac cycle as applied in many other existed publications. The advantage of this way is that the method developed can be employed to any echo videos without the need of extra ECG data. Smaller dataset is another factor. For 2D KAZE method, there are a lot of slides, while only 432 videos for 3D KAZE. In addition, we only sample one cardiac cycle, which leads to the lack of temporal information during detecting features. Therefore another future work is to align these data automatically without ECG data.

Comparison with other state-of-the-art, our method shows better performance in both four and eight viewpoints classification of echocardiogram videos.

8. Conclusion and future recommendations

This dissertation has described an echocardiogram video recognition method based on 3D KAZE detector and HOA descriptor. To conclude our work, we summarize our key contributions and discuss conclusions from our implementations in Section 8.1. Based on these conclusions, we then indicate a number of directions for future researches in the Section 8.2.

8.1 Conclusion

Our method demonstrates better performance for multiple viewpoints classification of echocardiogram video as shown in Table 7.4. After comparison, our approach shows promising results based on our dataset. In addition, it need to be improved to promote the echo classification precision in future work. The procedure in our research includes four stages: detecting and describing spatial-temporal features through an integrated echocardiogram video, encoding these features, finally classifying the videos into corresponding cardiac structure groups. According to these stages, we have proposed solutions to address these challenges with the following key contributions:

3D KAZE detector. We have presented a novel spatial-temporal feature detecting method, called 3D KAZE feature, which is an extension of 2D KAZE feature detector in the application of echocardiogram video classification. In practical implementation, we create a set of non-linear scale spaces and make blurring locally adaptive to the image data so that noise level can be reduced, while maintaining details and edges. It also highlights some boundary features and makes feature points more distinguishable because of the reducing of the influence of noises with little gradient changes. This detecting process can make the description of oriented histograms of image gradient (HOG) feasible for echocardiogram images, since the

scale invariant features (SIFT) as well as oriented histograms of image gradient description are ineffectual for echocardiogram images because of noisy gradients. In our evaluations as shown in Table 4.3, HOG descriptor delivers good performance in echocardiogram video classification.

HOA descriptor. Our second contribution is a local descriptor based on the histograms of acceleration fields where we evaluate for the task of echocardiogram video recognition. We analyze the cardiac acceleration field from the points of motion and myocardial stress, and then introduce it into feature description to reflect motion and the stress state of the heart. We divide the video into a set of sub-volumes and compute the histogram of acceleration for each sub-volume. The final descriptor is formed by concatenating all descriptors of sub-volumes. In this way, it is more reliable to recognize cardiac structures, as confirmed from our experimental results. In direct comparison with some current state-of-the-art descriptors, our approach shows better performance.

Feature representation. We compare and evaluate two popular encoding methods including Fisher vector and BoW for echocardiogram video representation. For BoW with k-means clustering, we adopt several nearest centers to cluster each feature instead of using one clustering center and spatial-temporal pyramid to highlight cardiac geometric structures. All of these modifications in our experiments improve the performance significantly comparing with the standard BoW. About Fisher vector, it shows better performance than modified BoW for echocardiogram video classification. Parameters in both methods are evaluated and optimized. In addition, we make an evaluation with the combination of FV with the Spatial Pyramid to reflect the spatial distribution information of features although the result does not indicate the improvement considerably. Considering the

efficiency and accuracy, we come to the conclusions for echocardiogram feature representations:

- For smaller number of Gaussian models applied to the creation of Fisher vector and code words size for BoW, Fisher vector outperforms BoW in echocardiogram video classification;
- With the sizes of both Gaussian models and code words increasing, modified BoW is more suitable than Fisher vector, because higher dimensional Fisher vector is time-consuming for the subsequent classification.

Classification. For multi-class SVM, we experimentally show that one-versus-all strategy outperforms one-versus-one strategy slightly on the classification accuracy, but without the superiority on the training efficiency. In the case of fewer classes and less training samples, such as our collection of datasets with eight classes and 432 videos, the former strategy can be applied to improve the performance of classification. As for the choice of kernel function in SVM, the experiment shows that linear kernel is less preferable than RBF (radial basis function) kernel on the accuracy and efficiency for the classification.

8.2 Future work

Although our method is focusing on the classification of echocardiogram videos, it could also applicable for the localization of cardiac structures, motion tracking and functional motion detection, etc. In the future, we would like to extend to the following applications and improvements:

Extension to more cardiac viewpoints classification. Our research has shown near 100% performance in 3 primary locations. In the future this accuracy will be

further improved for 8 viewpoints of echocardiogram video. Toward this end, we will collect more samples. In addition, with the availability of ECG data in the future collection, the alignment of time scale can be achieved, which can indicate the starting phase of a heartbeat cycle. It is very helpful for keeping all detection and description at the same stage of cardiac cycle, which consequently improves spatial and temporal correlations between different videos. So it can make further improvement on the accuracy, especially for even more viewpoint classification of echocardiogram videos.

Improvement on descriptors. Although HOA descriptor shows excellent results for echo video recognition, there are still rooms for improvement. One possible improvement is the evaluation of using state-of-the-art methods for calculation of velocity, which can improve the precision of acceleration field. In addition, motion model can also be described using local feature trajectories. Since trajectories follow local movements over time, they can offer more principled motion modeling. Accordingly, another interesting possibility is to combine local feature trajectories with HOA descriptor. Given a feature point, local feature trajectory can be obtained by using tracking technology, i.e. KLT tracker as applied in [144].

Promoting the efficiency. Fisher vector is applied as the default encoding method to represent all the features in our experiment. Additionally, we intend to add structural information to the Fisher vector by applying spatial-temporal pyramids, which improves the performance in the classification. One limitation of Fisher vector representation is that with the size of Gaussian models increasing, the vector dimension increases significantly, which leads to relatively expensive calculations in the following classification. In the future work, we intend to employ

more effective methods in encoding and classifying for improving the efficiency of classification.

SVM classification with different kernels. In our experiments, we only use linear and RBF kernels with default parameter settings to implement SVM classification. Although both of them show good performance in echocardiogram recognition, more kernels or parameter optimizing strategy can be applied to improve the classification performance.

Cardiac motion information detection. Built on feature detection, we can track some local feature points or local areas during the whole cardiac cycle and then extract local motion stages as well as motional correlation between structures. It can be combined with myocardial shapes and functional analysis to provide cardiac pathological diagnosis basis.

Additional future work direction can be referred to implementation of cardiac structure recognition combining with pathologic diagnostic information and extend our method in other applications, such as other medical images or video classifications, human action recognition, traffic surveillance system etc.

References

[1] http://www.bhf.org.uk

[2] http://www.cdc.gov/

[3] http://www.wpro.who.int/china/mediacentre/factsheets/cvd/en/

[4] Leticia Fernández-Friera, Ana García-Álvarez, Borja Ibáñez. Imagining the Future of Diagnostic Imaging. In: Rev Esp Cardiol, Vol. 66 No.02, 2013.

[5] Nagueh S.F., Appleto C.P., Gillebert T.G.. Recommendations for the evaluation of left ventricular diastolic function by echocardiography. In: Journal of the American Society of Echocardiography. Vol. 22, No. 2, pp.107-133, 2009.

[6] Waggoner AD, Bierig SM. Tissue Doppler imaging: a useful echocardiographic method for the cardiac sonographer to assess systolic and diastolic ventricular function. In: Journal of the American Society of Echocardiography, 14(12):1143–52, 2001.

[7] Stouylen A, Heimdal A, Bjornstad K, et al. Strain rate imaging by ultrasound in the diagnosis of regional dysfunction of the left ventricle. In: Echocardiography, 16(4):321–9, 1999.

[8] Geyer H, Caracciolo G, Abe H, et al. Assessment of myocardial mechanics using speckle tracking echocardiography: fundamentals and clinical applications. In: Journal of the American Society of Echocardiography, **23**:351–369, 2010.

[9] Teske AJ, De Boeck BW, Melman PG, Sieswerda GT, Doevendans PA, Cramer MJ. Echocardiographic quantification of myocardial function using tissue deformation imaging, a guide to image acquisition and analysis using tissue Doppler and speckle tracking. In: Cardiovasc Ultrasound 2007.

[10] Yaseen R.I., Ahmed M.K., Hamed W.A.. Assessment of abnormal LV myocardial deformation properties in obese patients by 2D based strain and strain rate imaging. In: The Egyptian Heart Journal, Volume 67 (3), pp. 183–191, September 2015.

[11] Landesberg G., Gilon D., Meroz Y., Georgieva M., Levin P.D., Goodman S. et al.. Diastolic dysfunction and mortality in severe sepsis and septic shock. In: European Heart Journal, 33:895-903, 2012.

[12] Lang RM, Badano LP, Mor-Avi V, Afilalo J, Armstrong A, Ernande L et al.. Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging. In: Eur Heart J Cardiovasc Imaging, 16:233-70, 2015.

[13] Patricia CM, Julian Breeze and Michael R. R.. The Top Ten Cases in Cardiac MRI and the Most Important Differential Diagnoses. In: Medical Imaging in Clinical Practice, Chapter 11, 2013. DOI: 10.5772/53444.

[14] Regueiro A, Garcia-Alvarez A, Sitges M, Ortiz-Perez JT, De Caralt MT, Pinazo MJ, et al.Myocardial involvement in Chagas disease: Insights from cardiac magnetic resonance. Int J Cardiol, 30:165(1):107-12, Apr. 2013.

[15] Rochitte CE, Oliveira PF, Andrade JM, Ianni BM, Parga JR, Avila LF, et al. Myocardial delayed enhancement by magnetic resonance imaging in patients with Chagas' disease: a marker of disease severity. In: J Am Coll Cardiol, 46:1553-8, 2005.

[16] Higgins C, Byrd B, Farmer D, Osaki L, Silverman N, Cheitlin M**:** Magnetic resonance imaging in patients with congenital heart disease. In: Circulation, 70:851-860, 1984.

[17] Razavi RS, Hill DL, Muthurangu V, Miquel ME, Taylor AM, Kozerke S, Baker EJ. Three-dimensional magnetic resonance imaging of congenital cardiac anomalies. In: Cardiol Young, 13: 461–465, 2003.

[18] Ahmed N, Carrick D, Layland J, Oldroyd KG, Berry C. The role of cardiac magnetic resonance imaging (MRI) in acute myocardial infarction (AMI). In: Heart Lung Circ, 22(4):243-55, Apr. 2013.

[19] Kim HW, Farzaneh-Far A, Kim RJ. Cardiovascular magnetic resonance in patients with myocardial infarction: current and emerging applications. In: J Am Coll Cardiol, 55(1):1-16 Dec 29 2009.

[20] Antman E, Braunwald E. Acute myocardial infarction. In: Heart disease: a textbook of cardiovascular medicine, 2000.

[21] Lo, S., Kwok, W.K. Acute myocardial infarction found by multi-detector computed tomography ordered for suspected aortic dissection. In: Hong Kong Med J. ,14:233–235, 2008.

[22] Raggi P, McLean D, Alexopoulos N. Coronary artery computed tomography. In: Zaret BL, Beller GA, eds. Clinical nuclear cardiology, 4th Edition. Elsevier; 2009.

[23]Nagueh S.F., Bierig S.M., Budoff M.J., *et al.* American Society of Echocardiography clinical recommendations for multimodality cardiovascular imaging of patients with hypertrophic cardiomyopathy: endorsed by the American Society of Nuclear Cardiology, Society for Cardiovascular Magnetic Resonance, and Society of Cardiovascular Computed Tomography. In: J Am Soc Echocardiogr, pp. 473–498, (24)2011.

[24] Vincent C. and Anahi P.: Basics of Ultrasound Imaging. In: Atlas of Ultrasound-Guided Procedures in Interventional Pain Management, 2011. DOI 10.1007/978-1-4419-1681-5_2.

[25] Atta E., Mohamed E. and Hosnia A.. Artificial Neural Networks in Medical Images for Diagnosis Heart Valve Diseases. International Journal of Computer Science Issues(IJCSI), Vol. 10, Issue 5, No 1,pp. 83-90, 2013.

[26]Sharma N., Ray A. K., Sharma S., Shukla K. K., Pradhan S. and Aggarwal L. M.. Segmentation and Classification of Medical Images Using Texture-primitive Features: Application of BAMtype artificial neural network. In: Journal of medical physics, 33(3), pp.119-126, 2008.

[27]Chykeyuk K., Clifton D. A., Alison Noble J.. Feature Extraction and Wall Motion Classification of 2D Stress Echocardiography with Relevance Vector Machines. In: IEEE, pp. 677-680, 2011.

[28] Beymer, D. and Syeda-mahmood, T.. Cardiac Disease Detection in Echocardiograms Using Spatio-temporal Statistical Models. In: Annual Conference of IEEE Engineering in Medicine and Biology Society (EMBS), 2008.

[29] Kumar, R., Wang, F., Beymer, D., Syeda-mahmood, T.. Cardiac Disease Detection from Echocardiogram using Edge Filtered Scale-Invariant Motion Features. In: IEEE Computer Society Workshop on Mathematical Methods in Biomedical Image Analysis, MMBIA, 2010.

[30] Takeshima, S., Matsuda, H., Yoshinaga, T., Masuda, K.. Development of Automatic Recognition Software of Left Ventricle by Time Series Processing Echocardiograms and Application to Disease Heart. In: Biomedical Engineering International Conference, BMEI, 2011.

[31] Agarwal D., Shriram K. S, Subramanian N.. Automatic view classification of echocardiograms using Histogram of Oriented Gradients. In: IEEE 10th International Symposium on Biomedical Imaging: From Nano to Macro San Francisco, CA, USA, pp 1368-1371, Apr. 7-11, 2013.

[32]Penatti O.A.B.,et al..Mid-level image representations for real-time heart view plane classification of echocardiograms. In: Comput. Biol. Med, 2015. http://dx.doi.org/10.1016/j.compbiomed.2015.08.004i.

[33] Ebadollahi, S., Chang, S.F., Wu, H.: Automatic view Recognition in Echocardiogram Videos Using Parts-based Representation. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 2-9, 2004.

[34]Aschkenasy S., Jansen C., Osterwalder R., Linka A., Unser M., Marsch S., and Hunziker P.. Unsupervised image classification of medical ultrasound data by multiresolution elastic registration. In: Ultrasound in Medicine and Biology, 32(7):1047–1054, July 2006.

[35]Balaji G. N., Subashini T. S. and Chidambaram N.. Cardiac View Classification using Speed Up Robust Features. In: Indian Journal of Science and Technology, Vol. 8(S7), pp. 1-5, 2015.

[36]Park J., Zhou S., Simopoulos C., Otsuki J., and Comaniciu D.. Automatic cardiac view classification of echocardiogram. In ICCV, pp. 1–8, 2007.

[37]Balaji G.N., Subashini T.S., Suresh A.: An Efficient View Classification of Echocardiogram Using Morphological operations. In: Journal of Theoretical and Applied Information Technology, Vol. 67 No.3 pp732-735, 2014..

[38] Kumar, R., Wang, F., Beymer, D., Syeda-mahmood, T.: Echocardiogram View Classification Using Edge Filtered Scale-Invariant Motion Features. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2009.

[39] D. Beymer, T. Syeda-Mahmood, and F. Wang. Exploiting spatio-temporal information for view recognition in cardiac echo videos. In: IEEE Computer Society Workshop on Mathematical Methods in Biomedical Image Analysis (MMBIA), pages 1–8, 2008.

[40] Kruger RP, Towns JR, Hall DL, et al. Automated radiographic diagnosis via feature extraction and classification of cardiac size and shape descriptors. In: IEEE Transactions on Biomedical Engineering, BME-19(No. 3):174–186, May 1972.

[41] Shiraishi J., Li Q., Suzuki K., Engelmann R., Doi K.. Computer-aided diagnostic scheme for the detection of lung nodules on chest radiographs: localized search method based on anatomical classification. In: Med Phys, 33:2642–2653, 2006.

[42] Vidya K.S., Ng E.Y.K, Acharya U.R., Chou S.M., Tan R.S., Ghista D.N.. Computer-aided diagnosis of Myocardial Infarction using ultrasound images with DWT, GLCM and HOS methods. In: Computers in Biology and Medicine. Volume 62 Issue C, pp. 86-93, July 2015.

[43]Hussein Z.R., Rahmat R.W., Abdullah L.N., Saripan M.I., Zamrin DM. Quantitative Detection of Left Ventricular Wall Motion Abnormality by Two-Dimensional Echocardiography. In: Computer and Information Science. Vol. 3, No. 2, May 2010.

[44] WFUMB 1997 World Federation for Ultrasound in Medicine and Biology News. In: Ultrasound Med. Biol., 1997.

[45]Wells P N T.. The medical applications of ultrasonics. In:Reports on Progress in Physics. 33:pp. 45–99, 1970.

[46]Ryding A.. Essential Echocardiography. In: Second Edition. Elsevier, 2013.

[47] Wikipedia. Echocardiography —Wikipedia, the free encyclopedia, 2015. URL https://en.wikipedia.org/wiki/Echocardiography .

[48] Lohr JL. and Sivanandam S.. Introduction to echocardiography. In Iaizzo PA (eds): Handbook of cardiac anatomy, physiology and devices, Chapter 18. Hummana Press Totowa, New Jersey2009; pp: 241.

[49] Harald Lutz, Rudolf Meudt. Manual of ultrasound. Berlin: Springer, 1984. doi:10.1007/978-3-642-69064-8.

[50] Vermilion, R.P.. Basic physical principles, in Echocardiography in Pediatric Heart Disease (Snider, R., ed.), Mosby-Year Book, St. Louis, MO, pp. 1-10 ,1997.

[51] http://www.web Harald Lutz. Basics of ultrasound.md.com/heart-disease/echocardiogram

[52] Rita N. Bakhru and William D. Schweickert. Intensive Care Ultrasound: I. Physics, Equipment, and Image Quality. In: Annals of the American Thoracic Society, Vol. 10, No. 5, pp. 540-548, 2013. doi: 10.1513/AnnalsATS.201306-191OT

[53] http://lsa.colorado.edu/essence/texts/heart.html. The Circulatory System. In: Part II: the heart and circulation of blood.

[54] Bijnens B, Cikes M, Butakoff C, Sitges M, Crispi F. Myocardial motion and deformation: What does it tell us and how does it relate to function? In: Fetal Diagn Ther 32: 5–16, 2012.

[55] http://www.tutorvista.com/content/biology/biology-iv/circulation-animals/heart-cardiac-cycle.php .Heart cycle time interval.

[56] Silky N., Madan L.. Adaptive Image Enhancement of Echocardiographic Images Using Automatic Roi. In: International Journal of Application or Innovation in Engineering & Management (IJAIEM). Volume 2, Issue 7, pp.148-154, July 2013.

[57] Bansal RC, Tajik AJ, Seward JB&Offord KP. Feasibility of detailed two-dimensional echocardiographic examination in adults: Prospective study of 200 patients. Mayo Clin Proc 55:291-308,1980.

[58] Abbott J.G, Thurstone F. L., "Acoustic speckle: Theory and experimental analysis," Ultrasonic Imaging, Vol 1, pp 303– 324, 1979.

[59]Bini A.A., Bhat M.S.. Despeckling low SNR, low contrast ultrasound images via anisotropic level set diffusion. In: Multidim Syst Sign Process, 25:41-65, 2014. DOI 10.1007/s11045-012-0184-5.

[60]Benzarti F. and Amiri H.. Speckle noise reduction in medical ultrasound images. In: International Journal of Computer Science, vol. 9, no. 2, pp. 187-194, 2012.

[61] Uddin M. S., Halder K. K., Tahtali M. et al. Speckle Reduction and Deblurring of Ultrasound Images Using Artificial Neural Network. In: Picture Coding Symposium (PCS), pp.105-108, 2015.

[62]Hu Peng S., HwanKim D., Seok-LyongLee N., KwanLim M.. Texture feature extraction based on a uniformity estimation method for local brightness and structure in chest CT images. In: Computers in Biology and Medicine (40), pp. 931-942, 2010.

[63]Lowe, D.. Distinctive image features from scale-invariant keypoints. In: Intl. J. of Computer Vision, 60, 91–110, 2004.

[64]Grauman K., Darrell T.. The pyramid matching kernel: Discriminative classification with sets of image features. In: Proceedings of the Tenth IEEE international Conference on Computer Vision, 2005.

[65]Mohammad Ali Maraci, Raffaele Napolitano, Aris Papageorghiou, Allison Noble J.. Object Classification in an Ultrasound Video Using LP-SIFT Features. In: Medical Computer Vision, LNCS 8848, pp.71-81,2014.

[66] Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.. SURF: Speeded up robust features. In: Computer Vision and Image Understanding 110, 346–359, 2008.

[67]Alcantarilla P. F., Bartoli A., and Davison A. J.. KAZE features. In Eur. Conf. on Computer Vision (ECCV), pages 214–227, 2012.

[68] Wei L., Yu Q., Martin L., and Xiaohong G.. The application of KAZE features to the classification echocardiogram videos. In: Multimodal Retrieval in the Medical Domain, volume 9059 of Lecture Notes in Computer Science, Vienna, Austria, Springer LNCS, 2015.

[69]Ni D., Chui Y., Qu Y., Yang X., Qin J., Wong T., Ho S., and Heng P.. Reconstruction of volumetric ultrasound panorama based on improved 3D SIFT. In: Computerized Medical Imaging and Graphics, 33(7):559–566, 2009.

[70] Gu X.Y. and Liu Y.S. . Point Cloud Coarse Registration Based on 3DSIFT Keypoint Detection. In: Journal of Computational Information Systems 10: 23, pp.10121–10128, 2014.

[71]Allaire S., Kim J., Breen S., Jaffray D., Pekar V.. Full orientation invariance and improved feature selectivity of 3D SIFT with application to medical image analysis. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 1-8, 2008.

[72]Yu T.-H., Woodford O. J., Cipolla R.. A Performance Evaluation of Volumetric 3D Interest Point Detectors. Int J Comput Vis, 2012. DOI 10.1007/s11263-012-0563-2.

[73]Willems G., Tuytelaars T., and Van Gool L.. An efficient dense and scale-invariant spatio-temporal interest point detector. In ECCV, 2008.

[74]Laptev I. and Lindeberg T.. Space-time interest points. In ICCV, 2003.

[75]Laptev I.. On space-time interest points. IJCV, 64:107-123, 2005.

[76]Harris C. and Stephens M.J.. A combined corner and edge detector. In Alvey Vision Conference, 1988.

[77] Dollár P., Rabaud V., Cottrell G., and Belongie S.. Behavior recognition via sparse spatio-temporal features. In: IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS), pp.65-72, 2005.

[78] Yu Q., Lianyi W., Chunyan W., Xiaohong G.. The Synergy of 3D SIFT and Sparse Codes for Classification of Viewpoints from Echocardiogram Videos. In: Medical Content-Based Retrieval for Clinical Decision Support. Springer; Berlin Heidelberg: 2013.

[79] Lucas B.D. and Kanade T.. An iterative image registration technique with an application to stereo vision. In: International Joint Conference on Artificial Intelligence, 1981.

[80]Lu W.C., Wang Y.C.F. and Chen C.S.. Learning dense optical-flow trajectory patterns for video object extraction. In: IEEE Advanced Video and Signal Based Surveillance Conference, 2010.

[81]Matikainen P., Hebert M., and Sukthankar R.. Trajectons: Action recognition through the motion analysis of tracked features. In: ICCV workshop on Video-oriented Object and Event Classification, 2009.

[82]Wang H., Klaser A., Schmid C., and Liu C.L.. Dense trajectories and motion boundary descriptors for action recognition. In: IJCV, 103(1):60–79, 2013.

[83]Farnebäck G.. Two-frame motion estimation based on polynomial expansion. In: Proceedings of the Scandinavian Conference on Image Analysis (SCIA), 2003.

[84]Laptev I. and Lindeberg T.. Local descriptors for spatio-temporal recognition. In SCVMA, 2004.

[85]Laptev I., Marsza lek M., Schmid C., and Rozenfeld B.. Learning realistic human actions from movies. In CVPR, 2008.

[86]Uijlings J., Duta I. C., Sangineto E., and Sebe N.. Video classification with densely extracted HOG/HOF/MBH features: An evaluation of the accuracy/computational efficiency trade-off. In: Int. J. Multimedia Inf. Retrieval, vol. 4, no. 1, pp. 33–44, Mar. 2015.

[87] Navneet D. and Bill T.. Histograms of oriented gradients for human detection. In: CVPR, volume 1, pp. 886-893. IEEE, 2005.

[88]Scovanner P., Ali S., and Shah M.. A 3-dimensional SIFT descriptor and its application to action recognition. In: MULTIMEDIA, 2007.

[89] Kläser, A., Marszałek, M., Schmid, C.. A spatio-temporal descriptor based on 3D-gradients. In: Proc. of BMVA British Machine Vision Conference, pp. 995–1004, 2008.

[90]Horn B. and Schunck B.. Determining optical flow. In: Artificial Intelligence, 17:185-203, 1981.

[91] Navneet D., Bill T., and Cordelia S.. Human detection using oriented histograms of ow and appearance. In: ECCV, Springer, pp. 428-441, 2006.

[92]Sivic J. and Zisserman A.. Video Google: A Text Retrieval Approach to Object Matching in Videos. In: Proceedings of International Conference on Computer Vision (ICCV), pp. 1470–1477, 2003.

[93] Sánchez J., Perronnin F., Mensink T., Verbeek J.. Image classification with the fisher vector: theory and practice. In: International Journal of Computer Vision: 1–24, 2013.

[94] Simonyan K., Parkhi OM., Vedaldi A., Zisserman A.. Fisher Vector Faces in the Wild. In: Proceedings of BMVC, 2013.

[95]Niebles J., Wang H., and Fei-Fei L.. Unsupervised learning of human action categories using spatial-temporal words. In BMVC, 2006.

[96]Wang H., Ullah M. M., Kläser A., Laptev I., Schmid C.. Evaluation of local spatio-temporal features for action recognition. In: British Machine Vision Conference (BMVC), pp. 124.1-124.11, 2009. DOI:10.5244/C.23.124.

[97] Piotr B. and Francois B.. Evaluation of Local Descriptors for Action Recognition in Videos. In: ICVS 2011, LNCS 6962, pp. 61–70, 2011.

[98] Yang, J., Yu, K., Gong, Y. and Huang, T.: Linear Spatial Pyramid Matching Using Sparse Coding for Image Classification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1794 – 1801 (2009)

[99] Lei B., Tan E-L, Chen S., Zhuo L., Li S., Ni D., et al. Automatic Recognition of Fetal Facial Standard Plane in Ultrasound Image via Fisher Vector. In: PLOS ONE 10(5): e0121838, 2015. DOI:10.1371/ journal.pone.0121838.

[100]Chatfield K., Lempitsky V., Vedaldi A., and Zisserman A.. The devil is in the details: an evaluation of recent feature encoding methods. In: BMVC, 2011.

[101]Oneata D., Verbeek J., and Schmid C.. Action and event recognition with Fisher vectors on a compact feature set. In: ICCV, 2013.

[102] Vadim K., Laptev I.. Efficient feature extraction, encoding and classification for action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2593-2600, 2014. DOI 10.1109/CVPR.2014.332

[103] Jaakkola T. and Haussler D.. Exploiting generative models in discriminative classifiers. In: NIPS, pp. 487–493, 1998.

[104] Florent P., S´anchez J., and Thomas M.. Improving the Fisher Kernel for Large-Scale Image Classification. In: ECCV, Part IV, LNCS 6314, pp. 143–156, 2010.

[105] Hui Wu, Dustin M. Bowers, Toan T. Huynh, and Richard Souvenir. Echocardiogram View Classification Using Low-level Features. In: IEEE 10th International Symposium on Biomedical Imaging: From Nano to Macro, San Francisco, CA, USA, pp 752-755, 2013.

[106]Zhou S., Park J., Georgescu B., Simopoulos J., Otsuki J., and Comaniciu D.. Image-based multiclass boosting and echocardiographic view classification. In CVPR, pages 1559–1565, 2006.

[107]Suganya R., and Rajaram S.. Content Based Image Retrieval of Ultrasound Liver Diseases Based on Hybrid Approach. American Journal of Applied Sciences 9 (6), pp. 938-945, 2009.

[108]Roy A., Sural S., Mukherjee J., and Majumdar A. K.. State-based modeling and object extraction from echocardiogram video. IEEE Transactions on Information Technology in Biomedicine, 12(3):366–376, 2008.

[109]Balaji G.N., Subashini T.S., Chidambaram N.. Automatic classification of Cardiac Views in Echocardiogram using Histogram and Statistical Features. In: Procedia Computer Science 46, pp.1569 – 1576, 2015.

[110]Rumelhart D., Hinton G., and Williams R.. Learning representations by back-propagating errors. In: *Nature*, vol. 323, pp. 533-536, 1986.

[111] Bay H, Tuytelaars T, Van Gool L. SURF: Speeded up robust features.In: Proceedings European Con-ference on Computer Vision, 110:407–17, 2006.

[112]Otey M., Bi J., Krishna S., Rao B., Stoeckel J., Han A. S., and Parthasarathy S.. Automatic view recognition for cardiac ultrasound images. In MICCAI: Intl Workshop on Computer Vision for Intravascular and Intracardiac Imaging, pages 187–194, 2006.

[113]Grauman K. and Darrell T.. Approximate correspondences in high dimensions. In NIPS, 2006.

[114] Ashley EA., Niebauer J.. Understanding the echocardiogram. In: Cardiology Explained, chapter 4. London, Remedica , 2004.

[115] Transthoracic echo. Echocardiographer.org. http://echocardiographer.org/TTE.html.

[116]Echocardiography in ICU. Stanford University. http://web.stanford.edu/group/ccm_echocardio/cgi-bin/mediawiki/index.php/Main_Page

[117] Zainudin M. N. S., Said M. M. and Ismail M. M.. Feature Extraction on Medical Image using 2D Gabor filter. In: Applied Mechanics and Materials, Vol. 52-54, pp. 2128-2132, 2011.

[118]Alvarez L., Guichard F., Lions P. L., and Morel J. M.. Axioms and fundamental equations of image processing. In: Arch. Rat. Mech. Anal., vol. 123, pp. 200–257, 1993.

[119] Perona P. and Malik J.. Scale-Space and Edge Detection Using Anisotropic Diffusion. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 12, No.7, 1990.

[120] Catté F., Lions P.L., Morel J.-M., Coll T.. Image selective smoothing and edge detection by nonlinear diffusion. In: SIAM J. Numer. Anal. 29, pp.182–193, 1992.

[121] Sarti A., Mikula K., Sgallari F.. Nonlinear Multiscale Analysis of Three-Dimensional Echocardiographic Sequences. In: IEEE Transactions on Medical Imaging, Vol. 18, No.6, 1999.

[122]Mikula K., Sgallari F.. Semi-implicit finite volume scheme for image processing in 3D cylindrical geometry.In: Journal of Computational and Applied Mathematics 161, pp.119–132, 2003.

[123]Tamura H., Mori S., and Yamawaki T.. Textural Features Corresponding to Visual Perception. In: IEEE Trans. Systems, Man, and Cybernetics, vol. 8, pp. 460-473, 1978

[124] Kuijper A.. The Deep Structure of Gaussian Scale Space Images. In: PhD thesis, Chapter 2, Utrecht University, 2002.

[125]Kačur J., Mikula K.. Solution of nonlinear diffusion appearing in image smoothing and edge detection. In: Appl. Num. Math. , Vol.17, pp.47-59, 1995.

[126] Drbl´ıkov´a, O., Mikula, K.. Convergence analysis of finite volume scheme for nonlinear tensor anisotropic diffusion in image processing. SIAM Journal on Numerical Analysis 46(1), pp.37–60, 2007.

[127] Welk, M., Steidl, G., Weickert, J.. Locally analytic schemes: A link between diffusion filtering and wavelet shrinkage. Applied and Computational Harmonic Analysis 24, pp.195–224, 2008.

[128]Cottet G.-H.. Diffusion approximation on neural networks and applications for image processing. F. Hodnett (Ed.), Proc. Sixth European Conf. on Mathematics in Industry, Teubner, Stuttgart, 3-9, 1992.

[129]Grewening S., Weickert J., and Bruhn A.. From box filtering to fast explicit diffusion. In: Proceedings of the DAGM Symposium on Pattern Recognition, pp. 533–542, 2010.

[130]Barash D., Israeli M., and Kimmel R. An accurate operator splitting scheme for nonlinear diffusion filtering. In: Scale-Space and Morphology in Computer Vision, pp. 281–289, 2001

[131]Weickert J., Romeny B. M. H. and Viergever M. A.. Efficient and reliable schemes for nonlinear diffusion filtering. In: IEEE Transactions on Image Processing 7(3):398-410, 1998.

[132] Lowe, D. Object recognition from local scale-invariant features. In: Proc. Seventh International Conference on Computer Vision, Corfu, Greece, pp. 1150–1157, 1999.

[133]Chatfield K., Lempitsky V., Vedaldi A., and Zisserman A.. The devil is in the details: an evaluation of recent feature encoding methods. In: BMVC, 2011.

[134]Oneata D., Verbeek J., and Schmid C.. Action and event recognition with Fisher vectors on a compact feature set. In: ICCV, 2013.

[135]Kantorov V., Laptev I.. Efficient feature extraction, encoding and classification for action recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2593-2600, 2014. DOI 10.1109/CVPR.2014.332

[136]Vedaldi A.and Fulkerson B.. VLFeat: An open and portable library of computer vision algorithms. http://www.vlfeat.org/, 2008.

[137] Chen S. and Ram N.. Large-scale web video event classification by use of fisher vectors. In: Applications of Computer Vision (WACV), IEEE Workshop on, pp. 15-22, 2013.

[138] Jonathan M., Mohamed C., and Robert S.. One against one" or "one against all: Which one is better for handwriting recognition with svms? In: Tenth International Workshop on Frontiers in Handwriting Recognition, 2006.

[139] Erin L. A., Robert E. S., and Yoram S.. Reducing multiclass to binary: A unifying approach for margin classifiers. In: The Journal of Machine Learning Research, 1:113-141, 2001.

[140] Chih-Chung C. and Chih-Jen L.. Libsvm: a library for support vector machines. In: ACM Transactions on Intelligent Systems and Technology, 2(3):27, 2011.

[141] Mihir J., Herve J., and Patrick B.. Better exploiting motion for better action recognition. In: CVPR, 2013.

[142] Ryan R. and Aldebaro K.. In defense of one-vs-all classification. In: The Journal of Machine Learning Research, 5:101-141, 2004.

[143]https://github.com/pablofdezalc/kaze

[144] Sun J., Mu Y., Yan S., and Cheong L.-F.. Activity recognition using dense long-duration trajectories. In: IEEE International Conference on Multimedia and Expo, 2010.

Appendix
Appendix 1: Common methods

A1.1 Harris3D feature detector

For Harris3D feature detecting, a spatial-temporal second-moment matrix is computed by using independent spatial and temporal scale values (σ_i, τ_i), a separable Gaussian smoothing function G (show in equation 4.2). It is shown as:

$$\mu(x, y, z; \quad _i, \tau_i) = G(x, y, z; s\sigma_i, s\tau_i) * \begin{pmatrix} {L_x}^2 & L_x L_y & L_x L_z \\ L_x L_y & {L_y}^2 & L_y L_z \\ L_x L_z & L_y L_z & {L_z}^2 \end{pmatrix} \tag{A1.1}$$

Where s is a parameter relating the integration scale for G to the local scales (σ_i, τ_i). The first-order derivatives of the video sequence v are defined as:

$$L_x(x, y, z; \sigma_i, \tau_i) = \partial_x(G * v) \tag{A1.2}$$

$$L_y(x, y, z; \sigma_i, \tau_i) = \partial_y(G * v) \tag{A1.3}$$

$$L_z(x, y, z; \sigma_i, \tau_i) = \partial_t(G * v) \tag{A1.4}$$

The final location of feature points are given by the maxima of

$$H = \det(\mu) - k\, trace^3(\mu), \quad H > 0 \tag{A1.5}$$

$$\det(\mu) = \lambda_1 \lambda_2 \lambda_3$$

$$trace = \lambda_1 + \lambda_2 + \lambda_3$$

Where λ_i is the significant eigenvalue of μ with $\lambda_1 \leq \lambda_2 \leq \lambda_3$. k is a coefficient. In this study, all the spatial and temporal scale levels are the same with part 4.3 and the default coefficient is $k = 0.0005$ (the similar setting with some researches [74] and [96]).

A1.2 Gabor feature detector

The Gabor detector, called Cuboid detector as well in some researches, applies a set of separable linear filters with 2D spatial Gaussian smooth kernel and 1D temporal Gabor filters. The response function is formulated as

$$R = (I * g * h_{ev})^2 + (I * g * h_{od})^2 \quad \text{(A1.6)}$$

Where $I(x, y, t)$ refers to the video sequence; $g(x, y; \sigma)$ is the 2D spatial Gaussian kernel with spatial scale σ, whereas $h_{ev}(t; \tau, \omega)$ and $h_{od}(t; \tau, \omega)$ defined as Eq.(3.22) are a quadrature (cosine and sine) pair of 1D temporal Gabor filters with temporal scale τ. The Gabor filters are defined as

$$h_{ev}(t; \tau, \omega) = -\cos(2\pi\omega t)e^{\frac{-t^2}{\tau^2}}$$

$$h_{od}(t; \tau, \omega) = -\sin(2\pi\omega t)e^{\frac{-t^2}{\tau^2}} \quad \text{(A1.7)}$$

Where $\omega = {}^{4}/_{\tau}$. The feature points are the local maxima of the response function R.

A1.3 Fisher vector

FV encoding assumes that descriptors are generated by a GMM (Gaussian Mixture Model) :

$$\mathcal{U}_\lambda(x) = \sum_{\mathrm{i}=1}^{\mathrm{K}} \omega_i \, \mu_i(x) \quad \text{(A1.8)}$$

Where $\lambda = \{\omega_i, \mu_i, \sum_i, \ i = 1 \ldots K\}$, $\omega_i, \mu_i, \sum_i$, are respectively the mixture weight, mean vector and covariance matrix of Gaussian $\mathcal{U}_{\mathrm{i}}$. Let $\gamma_{\mathrm{t}}(i)$ be the soft assignment of descriptor x_t to Gaussian *i*:

$$\gamma_{\mathrm{t}}(i) = \frac{\omega_i \mu_i(x_t)}{\sum_{j=1}^{K} \omega_j \mu_j(x_t)} \quad \text{(A1.9)}$$

The GMM model of *K* Gaussians, which can be viewed as 'codebook', is first obtained from a training set. Once the Gaussian model $(\omega_i, \mu_i, \sigma_i)$ is learned, the FV representation of the descriptors is given by the two parts [104]:

$$U_{\mu,i}^{X} = \frac{1}{T\sqrt{\omega_i}} \sum_{t=1}^{T} \gamma_t(i) \left(\frac{x_t - \mu_i}{\sigma_i} \right) \quad \text{(A1.10)}$$

$$V_{\sigma,i}^{X} = \frac{1}{T\sqrt{2\omega_i}} \sum_{t=1}^{T} \gamma_t(i) \left(\frac{(x_t - \mu_i)^2}{\sigma_i^{\,2}} - 1 \right) \quad \text{(A1.11)}$$

Where $U_{\mu,i}^{X}$ (resp. $V_{\mu,i}^{X}$) is the D-dimensional gradient with respect to the mean μ_i (resp. the standard deviation σ_i) of ith Gaussian with $i = 1 \ldots \ldots K$. The final representation is given by the concatenation of the two parts following the result of l_2-normalization.

A1.4 Bag-of-Words

Bag-of-Words, also called bag-of-features, is originally applied to document analysis. It can be applied to image/video classification, by treating visual features as words. In computer vision, a bag of visual features is a vector of occurrence counts of a dictionary of local image/video features. For the generation of visual dictionary, we apply either k-means or sparse coding on the set of training features. We randomly sample 80k features for training vocabulary. For results using k-means, features are assigned to their closest dictionary word using Euclidean distance. The resulting histograms of visual word occurrences are used as video sequence representations. For sparse coding, which models feature vector as a sparse linear combination of a set of basic elements by solving an optimization problem. The coefficients of each basic element are viewed as the representation of the corresponding input vector.

A1.5 Multiclass SVM

Classification is done with multiclass support vector machines. With regard to binary classification, an SVM aiming to learn a decision function based on the training dataset is defined:

$$f\left(\overline{F}\right)=\sum_{i=1}^{n} a_i k\left(\overline{F}_i, \overline{F}_j\right)+b \tag{A1.12}$$

Where $k(\overline{F_i}, \overline{F_j})$ is a kernel function including linear kernel and non-linear kernel. Different kernel functions have different results in classification. The trained videos are represented as $\{(\overline{F_i}, l_i)\}_{i=1}^{n}$, where $l_i \in \{1,2, \dots L\}$ denotes the class label of the trained video i. In our research, we study the influence on echocardiogram video classification from two kernels: linear kernel and Gaussian kernel (RBF)

For linear kernel:

$$k(\overline{F_i}, \overline{F_j}) = \overline{F_i}^T \overline{F_j} \tag{A1.13}$$

For Gaussian Radial Basis Function kernel:

$$k(\overline{F_i}, \overline{F_j}) = \exp(-\gamma \left\| \overline{F_i} - \overline{F_j} \right\|_2^2) \tag{A1.14}$$

For multiclass classification, we use the one-versus-all approach. In our implementation, we use the code provided by LIBSVM [140].

A1.6 Difference of Gaussian (DoG)

Similar to Laplace of Gaussian, the image is first smoothed by convolution with Gaussian kernel of certain width σ_1

$$G_{\sigma_1}(x, y) = \frac{1}{\sqrt{2\pi\sigma_1^2}} exp\left(-\frac{x^2+y^2}{2\sigma_1^2}\right) \tag{A1.15}$$

to get

$$g_1(x, y) = G_{\sigma_1}(x, y) * f(x, y) \tag{A1.16}$$

With a different width σ_2, a second smoothed image can be obtained:

$$g_2(x, y) = G_{\sigma_2}(x, y) * f(x, y) \tag{A1.17}$$

We can show that the difference of these two Gaussian smoothed images, called difference of Gaussian (DoG), can be used to detect edges in the image.

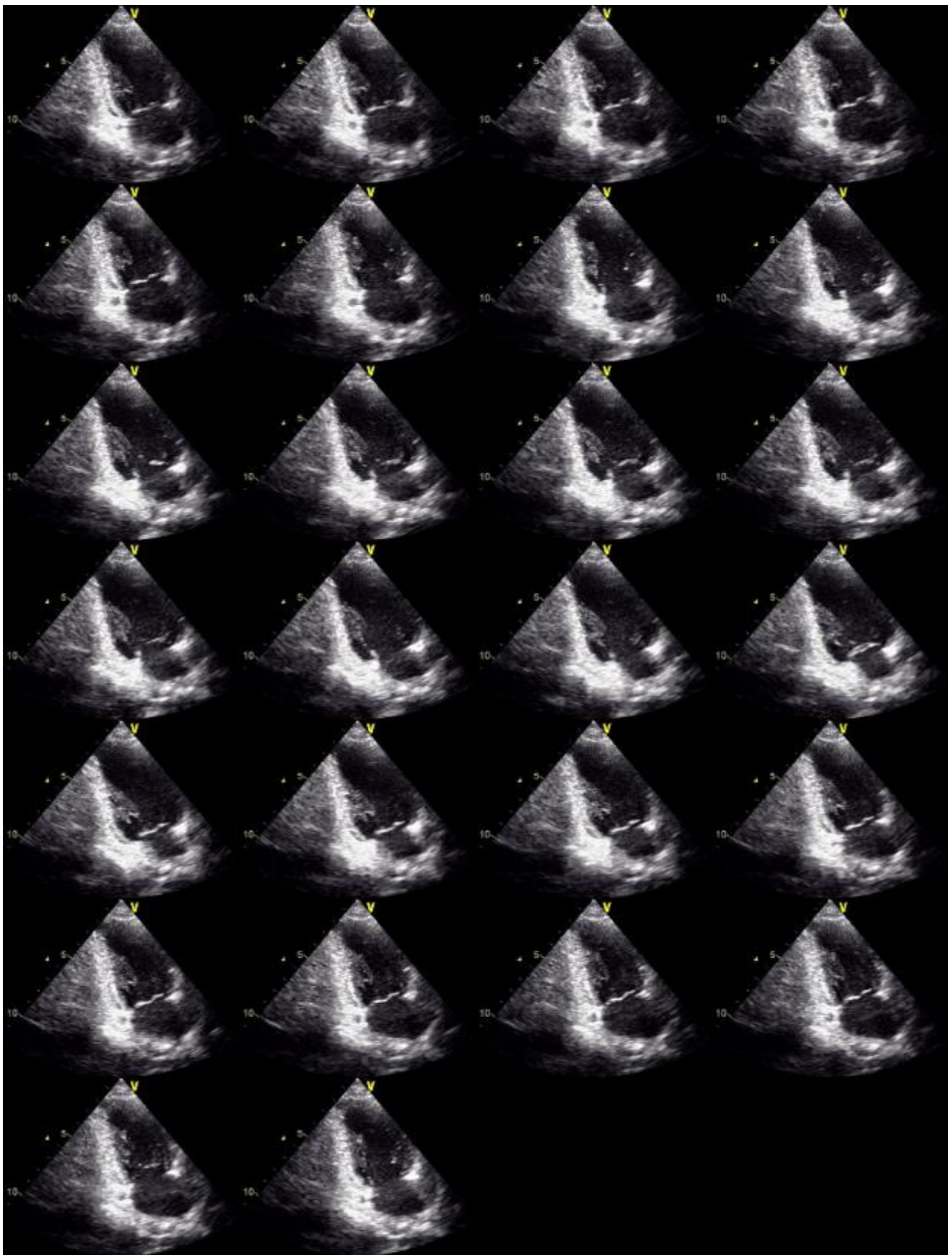
$$g_1(x,y) - g_2(x,y) = \left(G_{\sigma_1} - G_{\sigma_2}\right) * f(x,y) = DoG * f(x,y) \quad \text{(A1.18)}$$

The DoG as an operator or convolution kernel is defined as

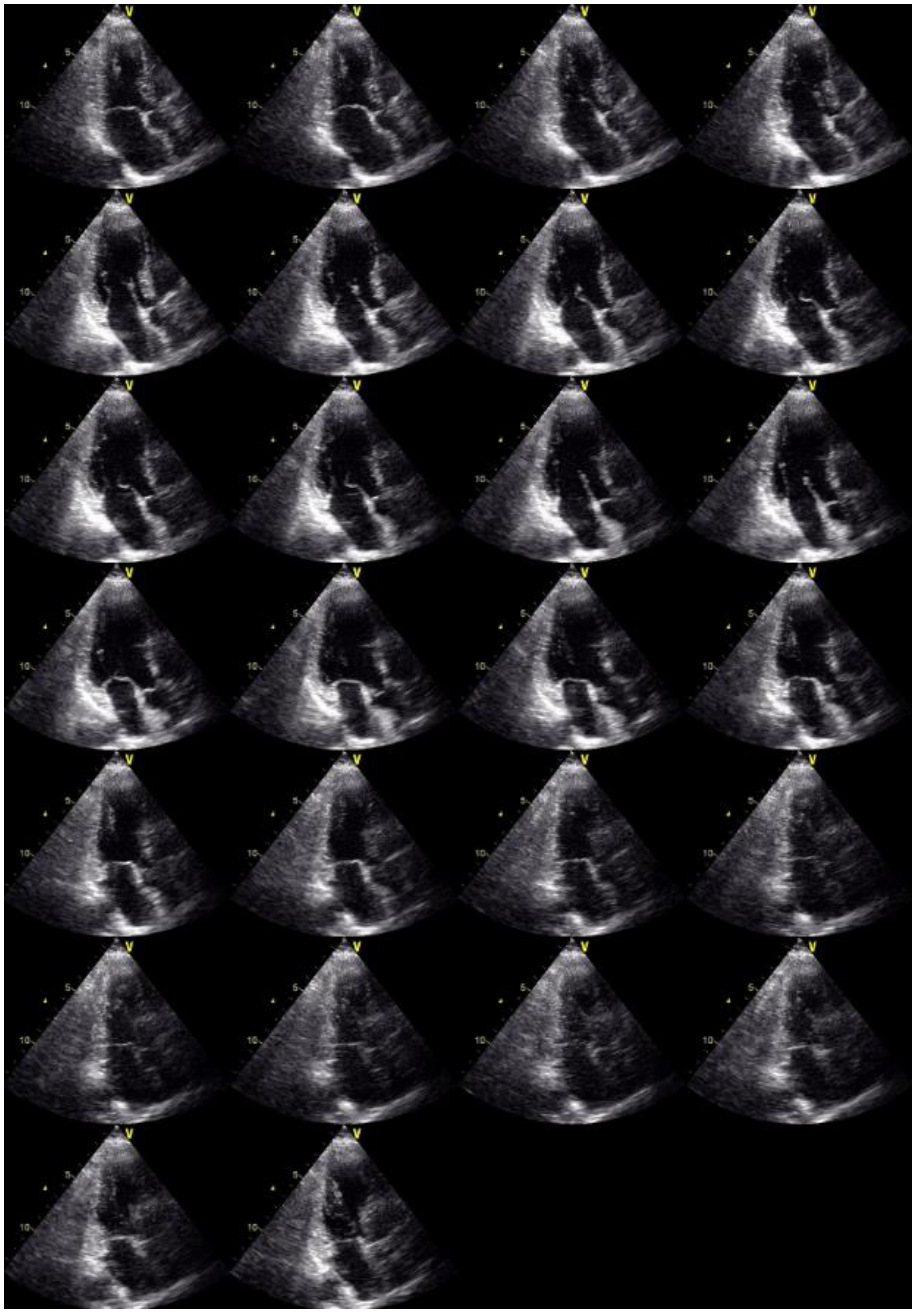
$$DoG \triangleq G_{\sigma_1} - G_{\sigma_2} = \frac{1}{2\pi}\left(\frac{1}{\sigma_1} e^{-(x^2+y^2/2\sigma_1^2)} - \frac{1}{\sigma_2} e^{-(x^2+y^2/2\sigma_2^2)}\right) \quad \text{(A1.19)}$$

Appendix 2: Video sample

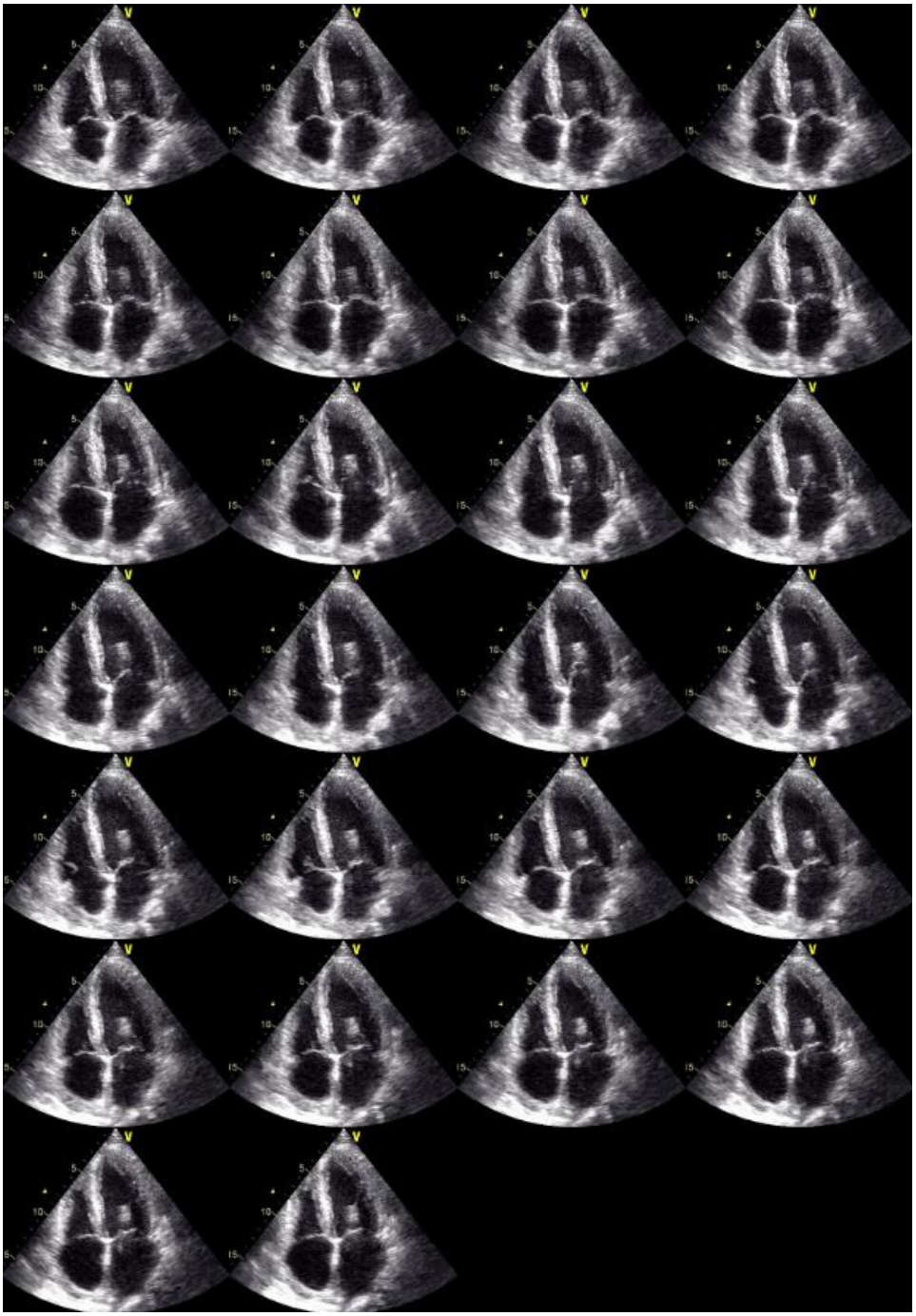
For A2C viewpoint:



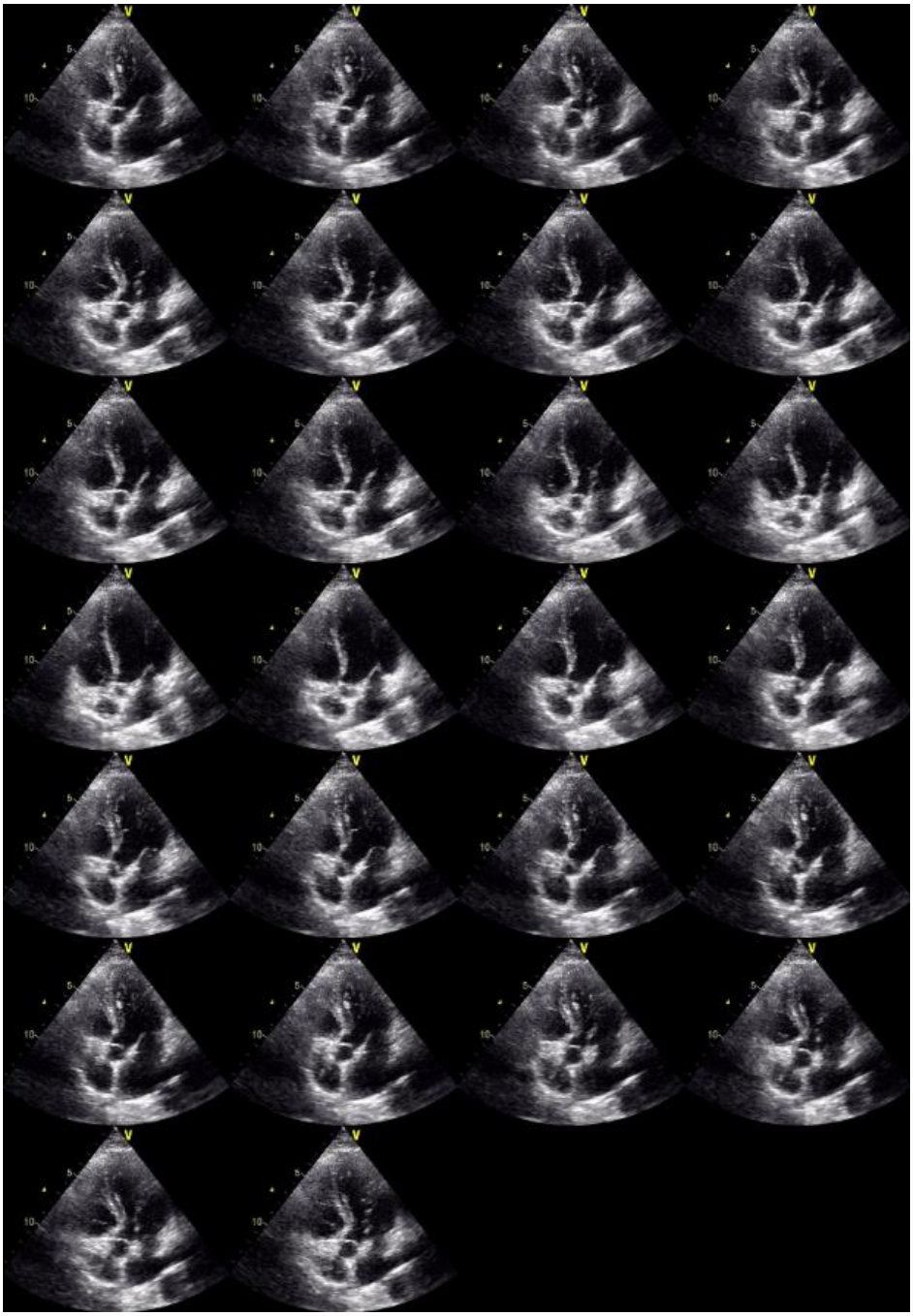
For A3C viewpoint:



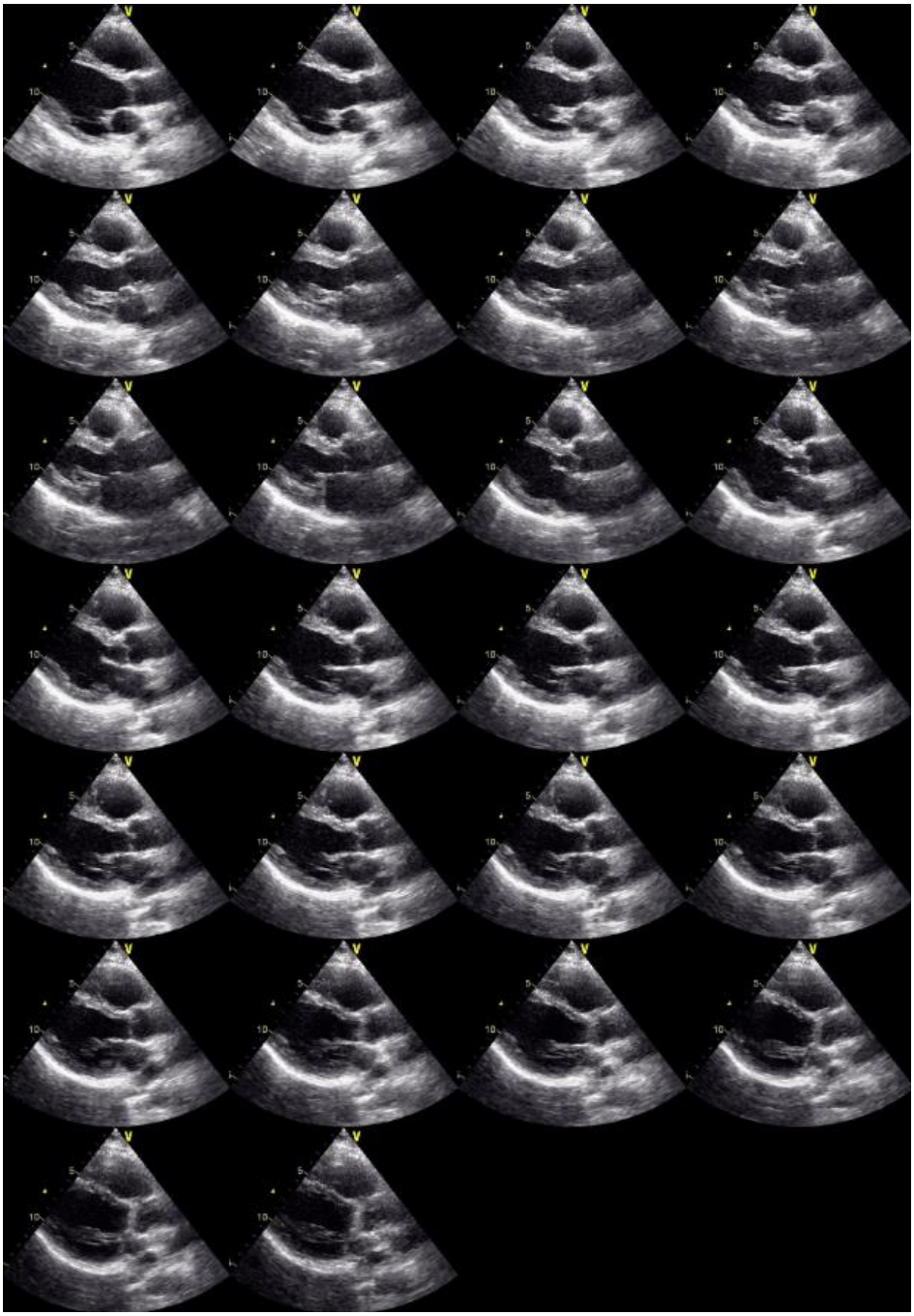
For A4C viewpoint:



For A5C viewpoint:

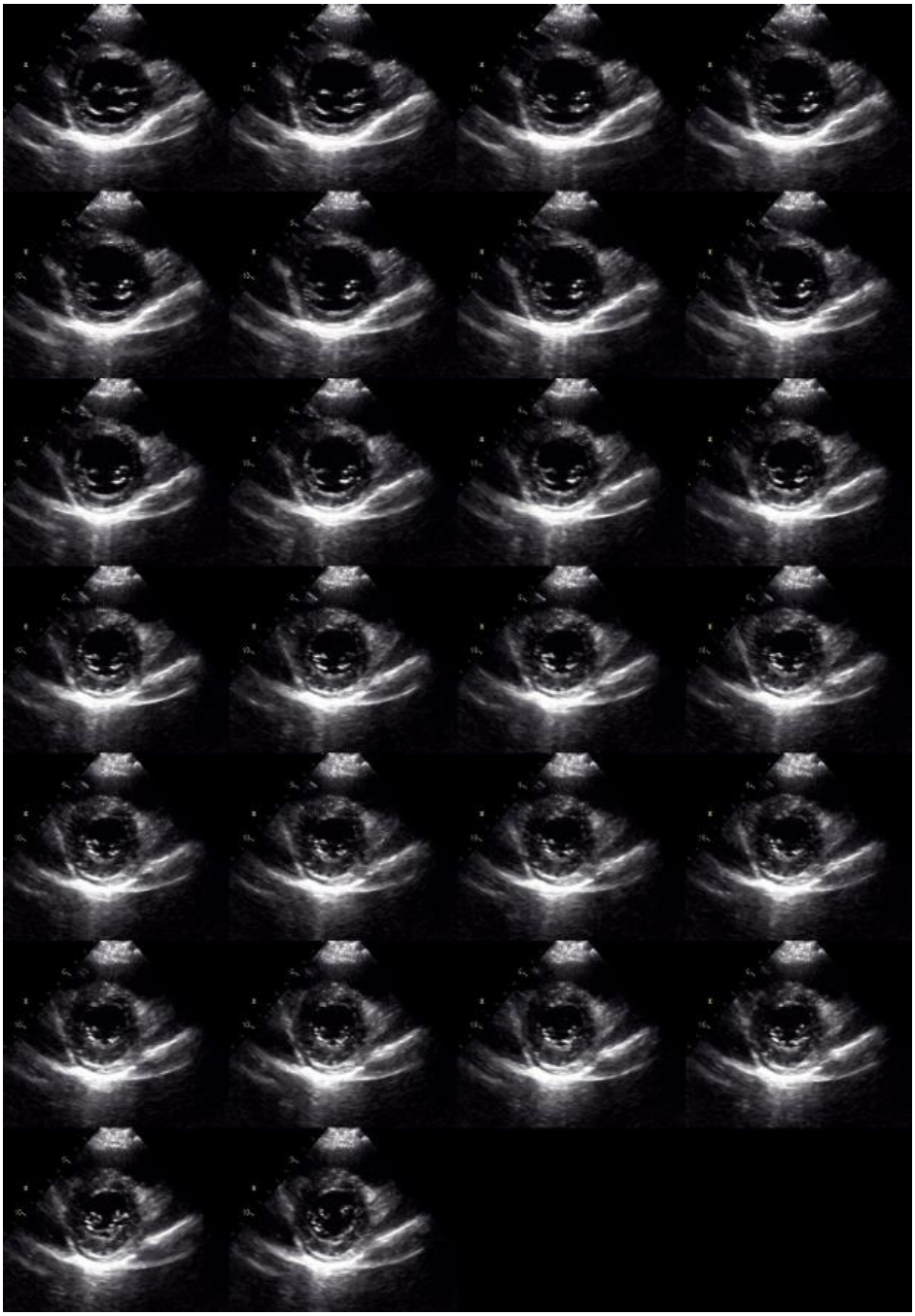


For PLA viewpoint:

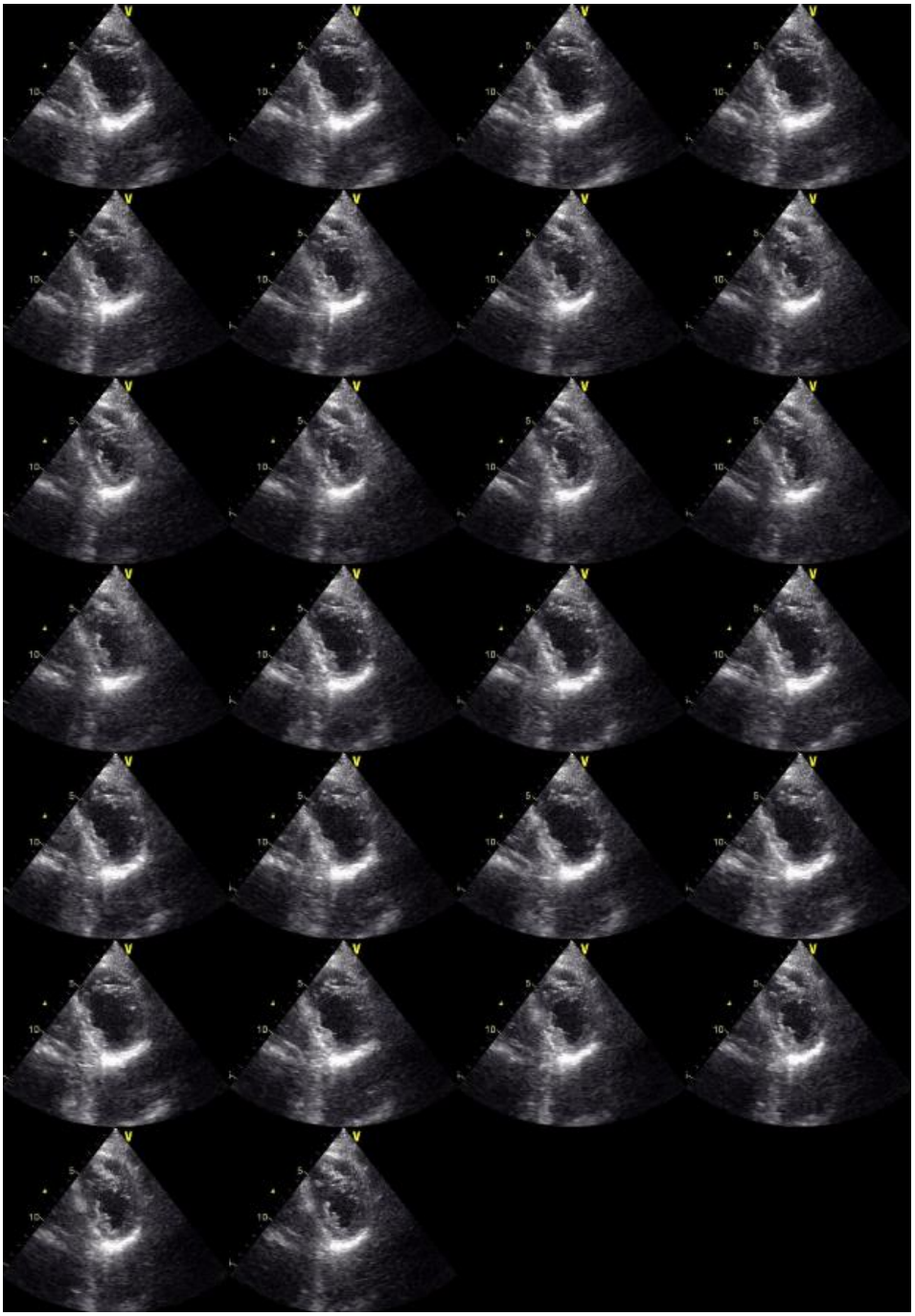


For PSAA viewpoint:

For PSAM viewpoint:



For PSAP viewpoint:



Appendix 3: Related publications

A. Gao, **W. Li**, M. Loomes, C. Lin, X. Gao. Cardiac Motion Reconstruction Using LKT Algorithm from 2D and 3D Echocardiography. In : the 2013 international conference on image processing, Computer Vision, and Pattern Recognition(IPCV 2013).

W. Li, Yu Qian, Martin Loomes, Xiaohong Gao. The application of KAZE features to the classification Echocardiogram Videos. In: Multimodal Retrieval in the Medical Domain (MRMD 2015).

Qiang Lin, **W. Li**. The Research of Sequential Images： Rebuilding of Gray (Position) ~time Function on Direction Lines and Their Applications. The Open Medical Informatics Journal. 2011.5.38-45.

L. Huang, X. zhang, **W. Li**. Dense Trajectories and DHOG for Classification of Viewpoints from Echocardiogram Videos. Computational and Mathematical Methods in Medicine Volume 2016 (2016), Article ID 9610192, 7 pages http://dx.doi.org/10.1155/2016/9610192